Simple Components, Correlated Components and an Application of Statistical Shape Analysis to Consumer and other Multivariate Data

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy

by

David Stanley Arnold

September 2010

Abstract

The interpretation of a principal component analysis can be complicated because the components are linear combinations of possibly many observed variables. A rotation of the principal components can improve the interpretation, however, there are usually still many small non-informative loadings, which taken together account for a significant proportion of the observed variation.

Presented is a new computationally efficient method to find simple components using similar criteria to principal components. Simple components are defined to have restricted weights that are proportional to the set of integers $\{0, \pm 1\}$. This choice ensures that no subjective decision is required as to whether a weight is important, and an individual weight is interpreted in a similar way to a correlation of one, minus one or zero with the component. The algorithm can find solutions for large problems in tractable time and can easily accommodate alternative criteria. An application is proposed that provides a simple component summary of a large data set.

When data is related to an orthogonal basis, these axes represent the maximum separation of information between axes. An approach is developed that finds orthogonal rotations of the principal components so that the sum or the sum of the squared covariance between a set of components is maximized. This approach can find a group of correlated components that explain a latent trait, and in addition explain different aspects of that trait. Another application is developed where an arbitrary configuration of points from a multidimensional scaling or similar method, can be displayed on a parallel coordinate plot so that the number of cross overs between the axes are minimized. This aids the identification of clusters and outliers.

In consumer research a respondent's perception is often driven by tacit knowledge, for example when making product comparisons. However, the traditional variable analogue scale may not capture this. A two dimensional response is proposed for a multiple product comparison. Principal shape analysis is developed to extract latent shape responses from the questions answered by the respondents. The analysis framework is coordinate free, and uses a scaled Euclidean distance matrix to represent a configuration of products, which can be considered a shape. A Euclidean distance matrix representation does not suffer from the problems associated with the use of shape coordinate systems.

Acknowledgements

Firstly, I would like to thank Dr Trevor Cox, without whose supervision, encouragement, and expertise, I would not have completed this thesis.

Secondly, Unilever Research for the opportunity to study and for providing the financial support and access to appropriate data sets.

My colleagues at Unilever Port Sunlight laboratory, for allowing the time and space to complete the work.

In particular I would like to thank Dr Jane Shaw for her encouragement, support, and suggestions, and help with the process of completing the thesis.
I would also like to thank Dr Tim Madden for his enthusiasm, advice and many useful discussions.

Dedication

To my sons, Chris and Jon

Contents

Chapter 1

Introduction

Multivariate data consists of more than one observation collected on each object or individual. There are often many variables which are usually correlated with each other and have an error structure that is more complex than in the case of a univariate data set. Also, the number of measured variables can be greater than the number of individual objects on which they are measured, for example data consisting of spectra or from some sensory product tests. Additionally, the measured variables may be of different types or on different scales. Consequently, the analysis, interpretation and quantification of uncertainty in such data is challenging. The high dimensional nature of most multivariate data makes it sparse and difficult to model statistically. Although work has been done on the estimation and hypothesis testing of population parameters, the methods cannot be used routinely and the majority of methods develop tools to explore and visualize the data and understand its structure. These methods can be considered exploratory. The following list highlights some of the questions that are commonly posed for multivariate data and gives examples of some methods. The list is not exhaustive, but is intended to give a flavour of the challenges and approaches commonly used.

1. Can the relationships be understood? This is concerned with obtaining the structure of the data. For example *graphical modelling*, *path analysis* and *structural equation modelling* all try to simplify the interdependencies between variables into a simple map to represent the true relationships. Hypothesis tests can be formulated to determine the likelihood of the relationships.

2. Can a simple set of variables be found to represent the data? An approach is to remove those variables that add little information, so called *variable selection*. Alternatively, a new set of variables are found that usually consist of linear sums of the original and simplify the correlation structure. *Principal component analysis* (PCA), *factor analysis* (FA) and *canonical correlation analysis* (CCA) are

examples. In doing this the error structure of models for the data is simplified.

3. Can observations and/or variables be classified or grouped? Clustering methods such as *hierarchical clustering* and *k-means clustering* group observations or variables based on some notion of distance. A *discriminant analysis* approaches the problem differently and finds boundaries in the multidimensional space in which the data sits, that separates observations into similar groups. Often the boundaries are formed using linear combinations of the variables, but non-linear boundaries can be fitted, for example using *support vector machines*.

4. Can a useful visualization be obtained? The previous three tasks in themselves produce a visualization of the data. In many cases the high number of variables is an inflated estimate of the true dimensionality of the data. A simple illustration is when the data sits on a straight line on a two dimensional graph. A *biplot* represents the relationships between the variables and the individual objects in two or three dimensions. These are particularly valuable when the true dimensionality of the data is of low order. The *generative topographic map* finds a flexible manifold embedded in the high dimensional space which can then be displayed in a small number of dimension, usually two, but preserves the ordering of the objects. Typically this is used to cluster observations in two dimensions. A different approach is a *parallel coordinate plot* which displays observations across the variables by representing them by vertical lines. Then relationships and outliers can be more easily identified.

5. Can the variables be regressed? *Multiple linear regression* and *multivariate analysis of variance* are methods where some assumptions are necessary regarding the dependencies and distribution of the variables and are multivariate generalizations of the univariate methods. However, algorithms exist to deal with independent and dependent sets that exhibit multiple correlation. For example, *partial least squares* finds linear sums for both the independent and dependent variable sets, and maximizes the covariance between the two sets.

The boundaries between the tasks are soft. There are also special types of multivariate data, such as ordered point sets used to model shape, and multivariate data collected over time. This thesis develops techniques that relate to principal component analysis, factor analysis and statistical shape analysis, with examples given from consumer data from the fast moving goods industry. The following sections of the introduction develop the ideas behind models that postulate a set of hidden variables called *latent variables* and later introduces statistical shape analysis.

Latent variables in the context of the models discussed in this thesis are constructed from weighted sums of the observed variables. In a *latent variable model* the observed

variables are termed *manifest variables* because the latent variables manifest their hidden relationships through them. However, in the first instance, *Principal Component Analysis* (PCA) is described. PCA finds weighted sums of the manifest variables which can then be considered latent variables. In the context of PCA these are termed *components*. PCA does not conform to the latent variable model framework primarily because it does not have an error model and explains all the variation in the data by its principal components. Later the principal component factor analysis model (FA) is explored where the model is adapted to be a form of the factor model and an error model is introduced. The purpose for outlining both PCA and FA is that this thesis develops ideas which simplify the structure of components and their interpretation or rotate factors to achieve a desired correlation, the main thrust of PCA and FA.

1.1 Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction method which seeks a small set of uncorrelated variables that explain as much of the variation present in the data as possible. As an example, consider a questionnaire which canvases respondents on their likes and dislikes of a hair conditioner. If seventy questions are posed, inevitably, a lot of information will be shared between the seventy questions and possibly a much smaller set of questions could capture the same information. PCA finds a linear combination of the observed variables that explains the maximum amount of variation possible. This is the first principal component. The next linear combination is then found that explains the next largest amount of variation but is uncorrelated with the first. This is the second principal component. This is repeated until a full set of uncorrelated components are found. It is hoped that the majority of variation explained by these components is captured by the first few components which then give a lower dimensional representation of the data. If the variance explained by this smaller set is large, for example 90% of the total variation, then the true dimensionality of the data is likely to be the cardinality of this smaller set e.g. three or four dimensional.

1.1.1 Derivation of Principal Components

Formally, principal components are obtained by finding the linear transformation of the random variables $\mathbf{x} = (x_1, \ldots, x_p)'$ to the principal components $\mathbf{y} = (y_1, \ldots, y_p)'$, where the y's are uncorrelated and labelled so that var $(y_1) \geq$ var $(y_2) \geq \ldots \geq$ var (y_p). The variable y_1 has the maximum variance possible subject to a length constraint, y_2 has the maximum variance possible and is uncorrelated with y_1, y_3 has the maximum variance possible and is uncorrelated with y_1 and y_2, and so forth. The derivation of principal

components can be found in many standard text books, for example Basilevsky (1994), Bartholomew and Knott (1987), Cox (2005), but is outlined here to link to later work in the thesis.

PCA transforms $\mathbf{x}$ to $\mathbf{y}$ such that

1. $y_j = a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{pj}x_p$ for $j = 1, \ldots, p$, $\mathbf{a}_j'\mathbf{a}_j = 1$, where $\mathbf{a}_j = (a_{1j} \ldots a_{pj})'$
2. $\text{corr}(y_j, y_k) = 0$ for $j \neq k$
3. y_j's are labelled so that variances are in descending order, i.e. $\text{var}\,(y_1) \geq \text{var}\,(y_2) \geq \ldots \geq \text{var}\,(y_p)$

Let the covariance matrix of $\mathbf{x}$ be $\mathbf{\Sigma}_X$. The first principal component is found by maximizing its variance. Let

$$V = \mathbf{a}_1'\mathbf{\Sigma}_X\mathbf{a}_1 - \lambda_1\left(\mathbf{a}_1'\mathbf{a}_1 - 1\right),$$

where λ is a Lagrange multiplier. Differentiation of V with respect to $\mathbf{a}_1$ gives

$$\frac{\partial V}{\partial \mathbf{a}_1} = 2\mathbf{\Sigma}_X\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = \mathbf{0},$$

and so

$$\left(\mathbf{\Sigma}_X - \lambda_1\mathbf{I}\right)\mathbf{a}_1 = \mathbf{0}.$$

This is a standard eigenvalue problem. The determinant of $\mathbf{\Sigma}_X - \lambda_1\mathbf{I}$ must be identically zero to obtain a solution. Hence λ_1 must be an eigenvalue of $\mathbf{\Sigma}_X$ and $\mathbf{a}_1$ its corresponding eigenvector. The variance of y_1 is

$$\begin{aligned}
\text{var}\,(y_1) &= \text{var}\left(\mathbf{a}_1'\mathbf{x}\right) \\
&= \mathbf{a}_1'\mathbf{\Sigma}_X\mathbf{a}_1 \\
&= \mathbf{a}_1'\lambda_1\mathbf{a}_1 \\
&= \lambda_1\mathbf{a}_1'\mathbf{a}_1 \\
&= \lambda_1.
\end{aligned}$$

Thus λ_1 is the largest eigenvalue of $\mathbf{\Sigma}_X$ and $\mathbf{a}_1$ its corresponding eigenvector. As $\mathbf{\Sigma}_X$ is positive definite all its eigenvalues must be real and greater than zero.

The next component is found in a similar way except that it is required to be uncorrelated with the first. This imposes an additional constraint on the maximization as $\text{cov}\,(y_1, y_2) = \mathbf{a}_1'\mathbf{\Sigma}_X\mathbf{a}_2$ and must be zero. Then the following is maximized

$$V = \mathbf{a}_2'\mathbf{\Sigma}_X\mathbf{a}_2 - \lambda_2\left(\mathbf{a}_2'\mathbf{a}_2 - 1\right) - \mu_{12}\mathbf{a}_2'\mathbf{\Sigma}_X\mathbf{a}_1, \tag{1.1}$$

where λ_2 and μ_{12} are Lagrange multipliers. After differentiation with respect to $\mathbf{a}_2$ and setting to zero,

$$\frac{\partial V}{\partial \mathbf{a}_2} = 2\boldsymbol{\Sigma}_X \mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 - \mu_{12}\boldsymbol{\Sigma}_X \mathbf{a}_1 = 0. \tag{1.2}$$

If this is pre-multiplied by $\mathbf{a}_1'$

$$2\mathbf{a}_1'\boldsymbol{\Sigma}_X \mathbf{a}_2 - 2\lambda_2 \mathbf{a}_1'\mathbf{a}_2 - \mu_{12}\mathbf{a}_1'\boldsymbol{\Sigma}_X \mathbf{a}_1 \quad = \quad 0.$$

Now, $\mathbf{a}_1'\boldsymbol{\Sigma}_X \mathbf{a}_2 = 0$ and $\mathbf{a}_1'\mathbf{a}_2 = 0$ by the orthogonality constraint, so $\mu_{12}\left(\mathbf{a}_1\boldsymbol{\Sigma}_X \mathbf{a}_1\right) = 0 \Rightarrow \mu_{12} = 0$. Substituting back into (1.2) gives

$$2\boldsymbol{\Sigma}_X \mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 = 0 \tag{1.3}$$

or

$$\left(\boldsymbol{\Sigma}_X - \lambda_2 \mathbf{I}\right)\mathbf{a}_2 = 0. \tag{1.4}$$

Again, the solution is an eigenvalue/eigenvector pair of $\boldsymbol{\Sigma}_X$ and must be the second largest eigenvalue λ_2 and its corresponding eigenvector. The variance of y_2 is λ_2. If this process is repeated the solutions are the p eigenvectors of $\boldsymbol{\Sigma}_X$ with corresponding eigenvalues being the variances. In some cases repeated eigenvalues may occur in which case there is not a unique eigenvector associated with each of these. In these circumstances choosing the eigenvectors to be orthogonal to those previously found ensures the previous arguments hold. Taking all the solutions together, let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)$ and then $\mathbf{y} = \mathbf{A}'\mathbf{x}$ and $\text{var}\left(\mathbf{y}\right) = \boldsymbol{\Delta}_Y$, where $\boldsymbol{\Delta}_Y = \text{diag}\left(\lambda_1 \ldots \lambda_p\right)$. In practice, the population covariance matrix $\boldsymbol{\Sigma}_X$ will not be known but can be estimated by the sample covariance matrix.

When the observed variables are not measured on the same scale it can become difficult to interpret the components especially if some of the variables have vastly larger variance, which then dominate the first principal component. The correlation matrix can be used instead of the covariance matrix. Unfortunately, there is not a simple relationship between the principal components of a covariance matrix and those of its correlation matrix.

The Spectral decomposition

Obtaining the principal components from multivariate data where the population covariance matrix $\boldsymbol{\Sigma}_X$ is estimated from the sample covariance matrix $\mathbf{S}_X$ can be viewed as a *Spectral Decomposition* of $\mathbf{S}_X$. The linear transformation $\mathbf{A}$ is the matrix which diagonalizes the symmetric positive definite covariance matrix $\mathbf{S}_X$ into the sum of its eigenvalue/eigenvector pairs.

$$\mathbf{S}_X = \mathbf{A}\,\boldsymbol{\Delta}_Y \mathbf{A}' = \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{a}_i' \,.$$

The Singular Value Decomposition

The singular value decomposition (SVD) is a convenient way to obtain the principal components. The SVD of the data matrix $\mathbf{X}$ is as follows,

$$\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}'$$

where $\mathbf{U}$ has columns consisting of the eigenvectors of $\mathbf{X}\mathbf{X}'$ and $\mathbf{V}$ the eigenvectors of $\mathbf{X}'\mathbf{X}$. Here $\mathbf{\Gamma}$ is a diagonal matrix of the shared singular values, and are the square root of the corresponding eigenvalues,

$$\mathbf{\Gamma} = \mathbf{\Delta}^{\frac{1}{2}}.$$

1.1.2 Biplots

Biplots are a set of visualisation techniques for multivariate data. If a technique like PCA explains most of the variation in a data set within the first two or three components, then biplots are a convenient and intuitive way to visualise the lower dimensional representation. The plot of the principal component scores is a graphical display for the observations and a plot of the coefficients of the first principal component against the second is a graphical display of the variables. A biplot displays both on the same axis. Biplots were introduced by Gabriel (1971) and an authoritative monograph on the subject is Gower and Hand (1996). Biplots can represent both continuous and categorical data. Points on the plot represent observations, and then axes are overlayed to represent the variables. The *classic biplot* is obtained by factorizing the data matrix into row (observations) $\mathbf{H}$ and column (variable) $\mathbf{G}$ matrices. If k is the rank of $\mathbf{X}$ then,

$$\mathbf{X}_{(N\times p)} = \mathbf{H}_{(N\times k)}\mathbf{G}_{(k\times p)}.$$

Approximating the data by approximating $\mathbf{H}$ and $\mathbf{G}$ by $N \times 2$ and $2 \times p$ matrices respectively,

$$\mathbf{X} \approx \mathbf{H}_2\mathbf{G}_2,$$

then the observations are represented by plotting $\mathbf{H}_2$ as points in a two dimensional space, and the variables are represented as axes on the same plot based on the p vectors of $\mathbf{G}_2$.

The matrices $\mathbf{H}$ and $\mathbf{G}$ are obtained from the singular value decomposition of $\mathbf{X}$, and then approximating $\mathbf{X}$ with the first two singular values, and introducing a parameter α,

$$\begin{aligned}
\mathbf{X} &= \mathbf{U}\mathbf{\Gamma}\mathbf{V}' \\
\mathbf{X} &\approx \mathbf{U}_2\mathbf{\Gamma}_2\mathbf{V}_2' \\
&\approx \left(\mathbf{U}_2\mathbf{\Gamma}_2^{\alpha}\right)\left(\mathbf{V}_2\mathbf{\Gamma}_2^{1-\alpha}\right)'
\end{aligned}$$

with $0 \leq \alpha \leq 1$. Different biplots are obtained by varying α. Then $\mathbf{U_2}\mathbf{\Gamma}_2^{\alpha} = \mathbf{H}_2$ is the $N \times 2$ matrix with each row representing an observation of $\mathbf{X}$ and $\mathbf{\Gamma}_2^{1-\alpha}\mathbf{V}_2' = \mathbf{G}_2$ is the $2 \times p$ vector with each column representing a variable. When $\alpha = 1$ this is a principal component biplot.

1.1.3 Example PCA

Much work has been done to frame PCA on the sample covariance matrix in an inferential setting. However, in this thesis the main purpose of a PCA is to reduce the dimensionality of a set of data and explore relationships. When a high dimensional data set is intrinsically of much lower dimension, plotting the first two or three principal component scores will give a straightforward visual representation of what the data looks like. This is because the first q principal components minimize the sum of the squared projection errors for each data point onto the subspace spanned by the q PCs. According to Jolliffe (2002) there are four areas to consider.

1. Which principal components are of interest
2. How many variables to keep
3. Are components considered sequentially or simultaneously
4. What is meant by 'approximating' the components

When considering which components are of interest, often the first q components are taken based on the proportion of variance explained in the data or a scree plot is used to see where the change in the variance explained flattens off. Alternatives, are to consider the eigenvalues with unequal variance as judged by a hypothesis testing procedure or by using a cross-validation process; again Jolliffe is an authoritative text. With an analysis on the sample covariance matrix, near zero eigenvalues may indicate the presence of linear dependencies between the variables which can then be dealt with (by removing a variable for instance). Linear dependencies and constant relationships between variables will show on the last principal components and often as large loadings on these. Therefore it is important also to look at the last principal components as a diagnostic aid.

The choice of how many variables to include is linked to how many components are considered. This is not trivial except for maybe a small number of variables. One way is to use a forward-backward stepwise approach, where subsets are evaluated based on how well they approximate the chosen subset of principal components (Cadima and Jolliffe, 2001).

PCA considers components sequentially, but in so doing is still optimal in terms of the variation explained. However, if components are desired which are more interpretable, then a sequential approach will probably not give a globally optimal solution for the chosen objective. This is linked to the final point of what is meant by approximating a set of components. However, in this thesis, approximating PCA is not the goal, but rather to find an interpretable set of components based on a defined simple set of loadings {-1, 0, 1 }.

The following example, of what might be termed a standard PCA, is taken from a sensory test where subjects were asked to access deodorant products by answering a questionnaire. There were 49 sensory questions scored on an ordinal scale, coded as 1 to 5, which represented a strong disagreement to a strong agreement. Additionally, there is a question scoring overall opinion on a 1 to 7 scale. Three test products were used, each subject assessed one of the products. There are 450 subjects giving 150 assessments per product. The attribute descriptors are listed in Table 1.2. One requirement of an analysis is to determine how the sensory data influences overall opinion. The correlation or regression of the principal components with overall opinion can be used to identify possible drivers. The data is taken as a whole in order to understand how the sensory questions relate to the overall opinion score. It is hoped that the products in the test will span the sensory space as measured by the questionnaire.

A PCA on the correlation matrix was used with overall opinion omitted, so that its relationship with the principal components of the sensory variables could be investigated later. The use of the correlation matrix simplifies the interpretation of the loadings, which are scaled as in Section 1.3.6 , and so also represent the actual correlation of the variables with the principal components.

Figure 1.1 is a scree plot, which shows the eigenvalues obtained from the analysis. There is a flattening off of the curve after the fifth eigenvalue. However, these five

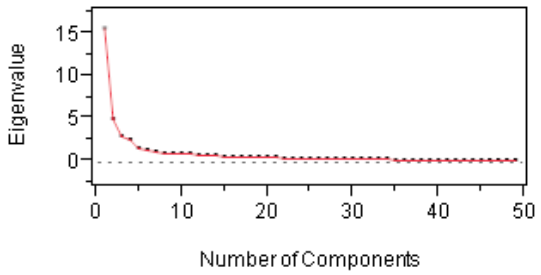


Figure 1.1: Scree Plot for the deodorant data

only explain 52% of the total variation, but this is typical of this type of sensory data. In fact, a good low dimensional representation of these data is not obtained with a PCA. However, these may still be useful to interrogate the data graphically. Table 1.2 show the loadings for the five principal components. An hierarchical cluster analysis on the five loading vectors helped to identify five groups of variables, which differentiate across the five vectors. The groups are listed in Table 1.1. For example the *Drying and Deposits* have high negative loadings on L1. L1 is capturing the contrast between positive and negative drivers of overall opinion. Indeed the subject scores for L1 are highly correlated with *overall opinion* which was left out of the PCA. Unfortunately, only this first and the second loading vectors are easy to label. It is useful to be able to identify labels for all the loading vectors. To make sense of the loading vectors, it is necessary to make subjective decisions regarding the importance of individual loadings, and to interpret across loading vectors. i.e. each loading vector cannot individually be labelled. The interpretation of the loading vectors and individual loading values (in this case also correlation), is not trivial. Cadima and Jolliffe (1995, 2001) explore

Group	Label
1	Use of the product applicator
2	Drying and deposits
3	Fragrance
4	Odour and wetness efficacy
5	Tactile properties

Table 1.1: Variable groups identified from the five principal component loading vectors

issues of interpreting the loadings and correlations. Apart from the first component where the variables with the largest absolute loadings are the most highly correlated, it is not possible to judge the correlation of a particular variable with subsequent components by purely examining the size of the loadings. When the correlation of a variable with a component is compared against the corresponding loadings, the size of a loading does not always give a good indication of the correlation of that variable with the principal component. In fact, the only time this is guaranteed to be the case is with the first principal component in which case the largest absolute loadings will have the largest correlations. Consequently, variables with high loadings may not be correlated as strongly as would appear and, conversely, variables with lower absolute loadings may be more highly correlated. This can also be the case when looking across component loadings where a given variable can have similar loadings but have very different correlations.

The aforementioned difficulties can make the interpretation of principal components difficult. Commonly, practitioners approximate principal components by ignoring loadings with small absolute values. Then, these truncated components are interpreted as

Code	Description	Group	L1	L2	L3	L4	L5
A1	rollball glided over skin	1	0.55	0.06	0.58	-0.03	0.24
A11	ball rolled freely in pack		0.49	-0.04	0.59	-0.07	0.25
A12	ball did not dry out		0.37	0.03	0.48	-0.10	0.24
A13	pack did not become messy		0.46	-0.06	0.41	-0.20	-0.13
A14	product did not leak out		0.39	-0.11	0.34	-0.23	-0.13
A16	easy of application		0.59	0.06	0.50	-0.07	0.10
A19	how smooth whilst applying		0.57	0.06	0.38	-0.01	0.05
A39	overall opinion packaging		0.41	0.09	0.29	-0.03	-0.08
A17	how product dosed from pack	2	-0.16	0.29	0.43	-0.03	-0.03
A43	felt wet during application		-0.64	0.39	0.28	-0.18	-0.17
A44	felt sticky whilst drying		-0.69	0.38	0.17	-0.11	-0.11
A45	left visible deposits		-0.60	-0.08	-0.03	0.22	0.37
A46	cold on application		-0.54	0.07	-0.01	0.03	0.13
A47	marked clothes		-0.55	0.04	0.04	0.33	0.42
A48	waited longer than usual- drying		-0.68	0.38	0.26	-0.16	-0.06
A49	felt greasy		-0.70	0.30	0.02	0.04	0.13
A33	overall opinion fragrance	3	0.42	0.30	0.09	0.60	-0.26
A34	strength fragrance-immediately		0.08	0.42	-0.03	0.51	-0.23
A35	strength fragrance- end of day		0.29	0.55	-0.06	0.44	-0.09
A6	had a pleasant fragrance		0.38	0.29	0.14	0.57	-0.33
A7	fragrance lasted long enough for me		0.46	0.53	-0.10	0.46	-0.15
A10	kept me fresh all day	4	0.67	0.50	-0.23	-0.12	0.18
A28	overall opinion - effective		0.80	0.28	-0.12	0.00	0.12
A29	notice any perspiration		0.45	0.40	-0.30	-0.31	0.14
A30	overall how effective keeping you dry		0.64	0.43	-0.27	-0.29	0.13
A31	notice any odour		0.33	0.48	-0.16	-0.18	0.15
A32	how effective keeping free from odour		0.56	0.54	-0.15	-0.19	0.11
A36	notice visible deposits - skin		0.39	0.00	-0.10	-0.37	-0.52
A37	notice deposits on clothes		0.34	0.01	-0.13	-0.35	-0.51
A38	how easy to wash off skin		0.22	-0.15	0.06	0.02	-0.24
A40	any irritation		0.21	-0.05	0.03	0.09	0.02
A41	any trapping of underarm hair		0.23	0.00	0.15	-0.12	-0.16
A42	how often applied rollon		-0.02	0.05	-0.05	-0.05	0.07
A8	gave me daylong protection - BO		0.61	0.53	-0.21	-0.15	0.19
A9	gave me daylong protection- wetness		0.63	0.43	-0.29	-0.26	0.21
A15	easy to apply the right amount	5	0.52	-0.18	0.27	0.06	0.12
A18	ease of applying right amount		0.55	-0.08	0.14	0.07	0.14
A2	felt fresh whilst applying		0.61	0.08	0.23	0.23	0.03
A20	how sticky whilst applying		0.69	-0.36	-0.11	0.10	0.01
A21	how greasy whilst applying		0.61	-0.25	0.09	-0.10	-0.06
A22	how wet whilst applying		0.59	-0.38	-0.33	0.15	0.10
A23	how cold whilst applying		0.50	-0.14	-0.08	-0.05	-0.14
A24	how sticky immediately after application		0.67	-0.38	-0.14	0.10	0.03
A25	speed of drying		0.64	-0.37	-0.30	0.13	0.19
A26	how sticky whilst wearing		0.65	-0.11	-0.04	-0.03	-0.17
A27	how greasy whilst wearing		0.62	-0.07	0.11	-0.16	-0.16
A3	felt smooth whilst applying		0.63	-0.03	0.32	0.17	0.14
A4	dried quickly		0.70	-0.40	-0.30	0.15	0.08
A5	left underarm soft and smooth		0.65	-0.01	-0.02	0.19	-0.03

Table 1.2: The Sensory Attribute Descriptions for the deodorancy example

weighted averages or contrasts of the remaining loadings. Firstly, together a large number of small loadings will account for a significant proportion of the observed variation and, secondly, approximating components in this way can lead to misinterpretation due to these truncated components poorly approximating the original components. A different subset of the same size may provide a better approximation. Also, correlations between individual variables and principal components are not appropriate, except when judging the adequacy of single-variable approximations. In fact, except for a single variable approximation, finding a subset of variables that closely approximates the original requires a combinatorial search. Cadima and Jolliffe recommend measures, and a search strategy to find the best subsets for moderate size problems.

The following flawed approach is often taken in practice. Considering the deodorant data and ignoring the small loadings, general patterns may be observed between the component loadings and the variable groups and these are shown in Table 1.3.

Group	Description	L1	L2	L3	L4	L5
1	Use of the product applicator	+	0	+	-	+/-
2	Drying and deposits	-	+	+	+/-	+
3	Fragrance	+	+	0	+	-
4	Odour and wetness efficacy	+	+	-	-	+/-
5	Tactile properties	+	-	+/-	+/-	+/-

Table 1.3: Group structure across the loadings by reducing the weights to the sign of the majority weight in that group on each component

If the panellist scores are plotted against each other then this can highlight unusual subjects. Figure 1.2 show the scores. To illustrate, the triangles are subjects whose scores on the first principal component are low, looking at Table 1.3, this may indicate that this subject scored higher on the group two variables and had low overall opinion. The cross, is a subject who has a low score on component four. Most of the negative loadings on this component are associated with efficacy and also use of the applicator. So this subject is scoring high on efficacy and applicator use, but has a moderate to low overall opinion of the product they tested. The square, is a subject who in addition to scoring low on component one also scored low on the third. Group three constitute the bulk of the negative loadings on this component, indicating that this subject has scored high on efficacy, but did not like the product overall. As only 52% of the variation is explained by the five chosen components, the differences between the labelled groups of variables are blurred, in particular group two and five. These may differentiate better with more components, but the interpretation will become more difficult and potentially more misleading as more loadings are ignored. Later, in Section 1.3.8 a rotation of the principal component axes is used to try to separate the loading vectors into independent latent variables. However, many small loadings are still present. Chapter 2 develops the ideas behind finding simple components, which do not require subjective decisions

to be made on whether a particular loading is important or not, and this can lead to more interpretable components.

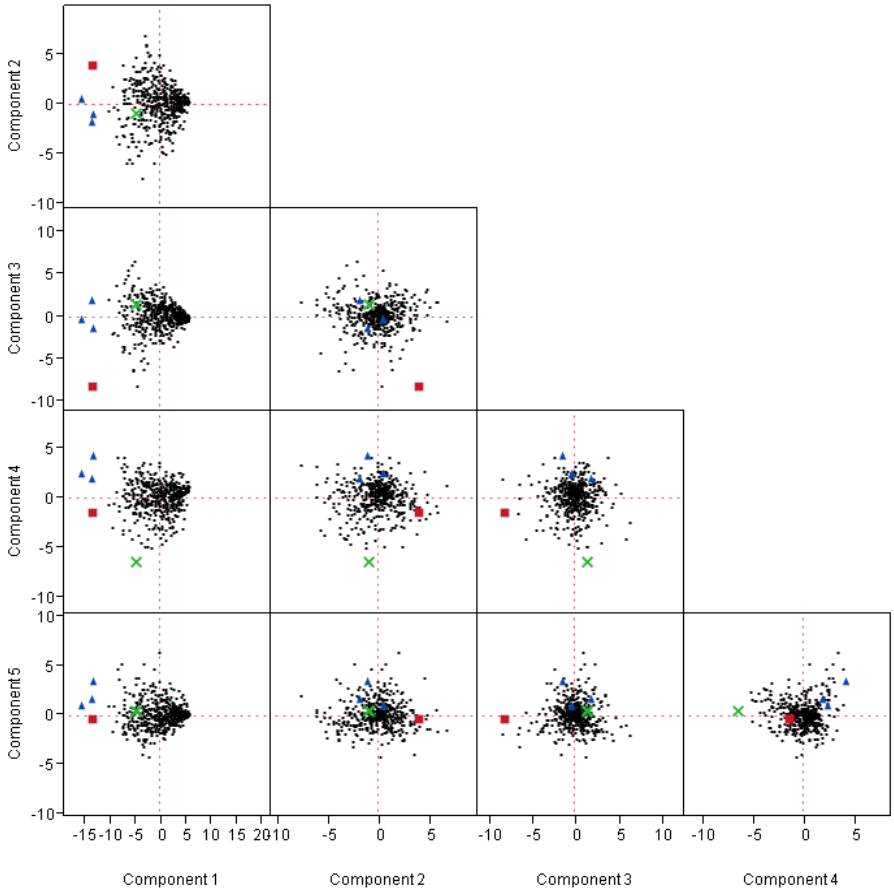


Figure 1.2: A Scatterplot matrix of the principal component scores.

Figure 1.3 shows a principal component biplot for the loadings and scores. Note, that the scores are rescaled to be between -1 and 1. The first two components only capture 37% of the variation in the data, this is reflected by many variables which have short vectors when projected onto this two dimensional representation. As component one explains 28% of the variation, it clearly separates low and high overall opinion. Drying and deposits represent opposite ends of the same axis in two dimensions. Fragrance and efficacy are correlated with each other. This is not surprising, as one could imagine that fragrance would influence the perception of odour efficacy.

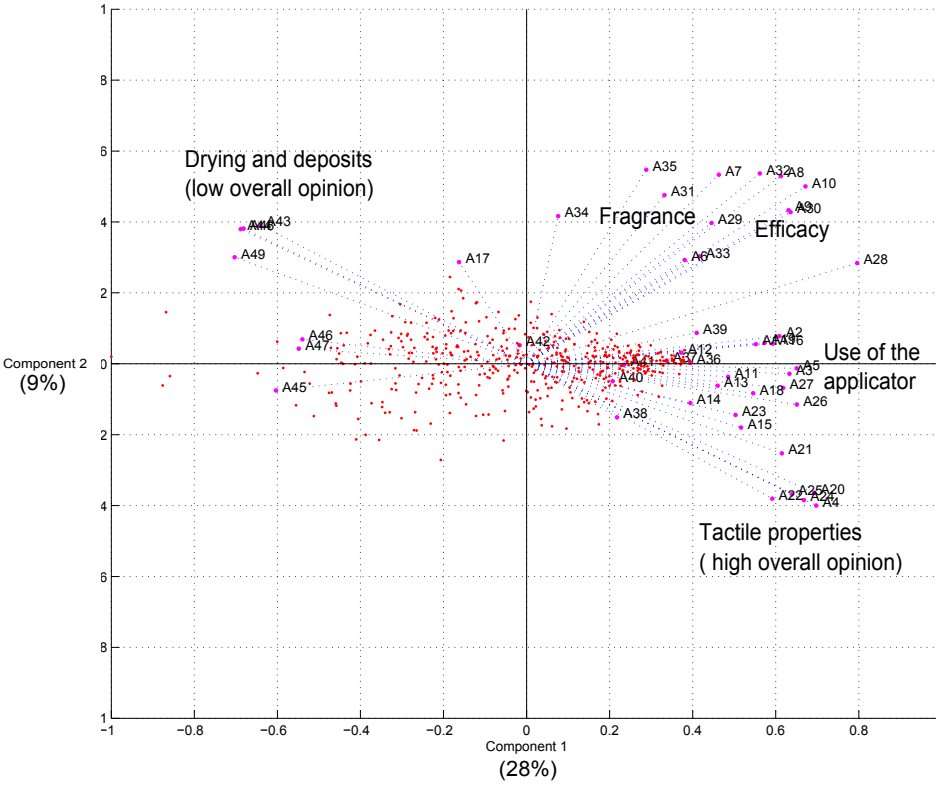


Figure 1.3: A biplot showing the sensory variables on the first two principal components. Some of variable clusters identified differentiate well on the first two principal components, however, clearly a higher dimensional representation is required as only 37% of the variation is captured by these two components.

Prin1	0.70
Prin2	0.17
Prin3	-0.10
Prin4	0.21
Prin5	0.08

Table 1.4: The correlation of Overall Opinion with the principal component scores

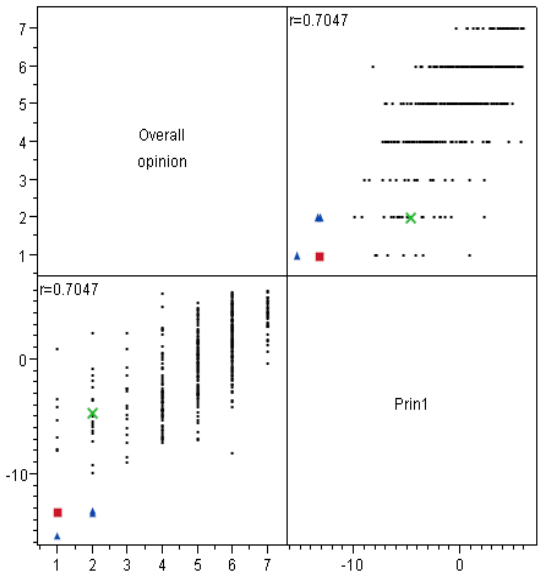


Figure 1.4: A plot showing the overall opinion scores against the first principal component scores for the deodorant data

1.2 Multidimensional Scaling

PCA can give a low dimensional map of objects which sit in high dimensional space. However, there is a broader set of methods called *Multidimensional Scaling (MDS)* which take a set of dissimilarities (or similarities) between objects and then find a configuration of points in a low dimensional space where each point represents one of the objects, and is such that distances between pairs of points in the space match the dissimilarities between the corresponding pairs of objects or the rank order of distances correspond to the rank order of dissimilarities; *matching* being that which is best achievable in some sense. An authoritative monograph on the subject is Cox and Cox (2000).

MDS methods can be split into *metric* and *non-metric* scaling. For metric scaling, the dissimilarities $\{\delta_{rs}\}$ between pairs of n objects are represented directly by distances $\{d_{rs}\}$ between pairs of corresponding points in the configuration.

One type of metric scaling is *Classical Scaling*. Very briefly, let $\mathbf{G} = \left[-\frac{1}{2}d_{rs}^2\right]$ and this is centred to give $\mathbf{B}$, i.e. $\mathbf{B} = \mathbf{HGH}$ where $\mathbf{H}$ is the centring matrix. Now $\mathbf{B} = \mathbf{XX}'$ is the inner product matrix (corrected for translation so that its centroid is at the origin) and is positive semi-definite with p non-zero eigenvalues and $n - p$ zero eigenvalues. The spectral decomposition of $\mathbf{B}$ (see 1.1.1) is $\mathbf{B} = \mathbf{U\Delta U}'$ and taking the p non-zero eigenvalues and corresponding eigenvectors so that

$$\mathbf{B} \approxeq \mathbf{U}_p\mathbf{\Delta}_p\mathbf{U}_p'$$

then the coordinates are recovered from

$$\mathbf{X} = \mathbf{U}_p \boldsymbol{\Delta}_p^{\frac{1}{2}}.$$

It can be shown that if Euclidean distance is used as a measure of dissimilarity then classical scaling is equivalent to PCA. In this case the points on the map are projected onto the principal axes.

In general, metric scaling will often minimize a loss function of the form

$$\frac{\sum_{\mathrm{r<s}} w_{rs} \left(d_{rs} - \delta_{rs}\right)^2}{\sum_{\mathrm{r<s}} \delta_{rs}},$$

sometimes called a *strain*. An example is a Sammon map (Sammon, 1969).

In a non-metric scaling, the magnitude of the distances between pairs of points no longer approximate the corresponding magnitude of the original dissimilarities. Instead, the rank orders are matched as well as possible. One approach to obtain a representation of the rank order is the minimization of Kruskal's (1964) loss function of the form

$$S = \sqrt{\frac{S^*}{T^*}},$$

where $S^* = \sum_{\mathrm{r<s}}^{n} \left(d_{rs} - \hat{d}_{rs}\right)^2$ and $T^* = \sum_{\mathrm{r<s}} d_{rs}^2$ is a normalizing term. S is termed the *Stress* function and $\hat{d}_{rs}$ are called disparities. The set of disparities $\{\hat{d}_{rs}\}$ are obtained by fitting a monotone least squares regression of the $\{d_{rs}\}$ on the dissimilarities $\{\delta_{rs}\}$. An iterative algorithm is used to minimize the stress.

A useful application of later work in the thesis is the convenient display of three dimensional or higher MDS configurations. In the case of non-metric scaling the axes are arbitrary and so can be rotated. If the configuration is rotated to maximize the correlation between object coordinates, then this can aid the search for clusters and outliers. In the case of classical scaling the points are projected onto the principal axes and as such have meaning relevant to these. However, rotation to a new set can still be useful to identify clusters and outliers. An application discussed in this thesis is to minimize the number of cross overs in a parallel coordinate plot given a configuration of points obtained from a MDS method. This could be any configuration that can be rotated (See Chapter 3).

1.3 The Latent Variable Model (LVM)

The latent variable model framework is first outlined in general. A good introductory text is Everitt (1984). Given vectors of observed variables $\mathbf{x}' = [x_1, \ldots, x_p]$ and latent

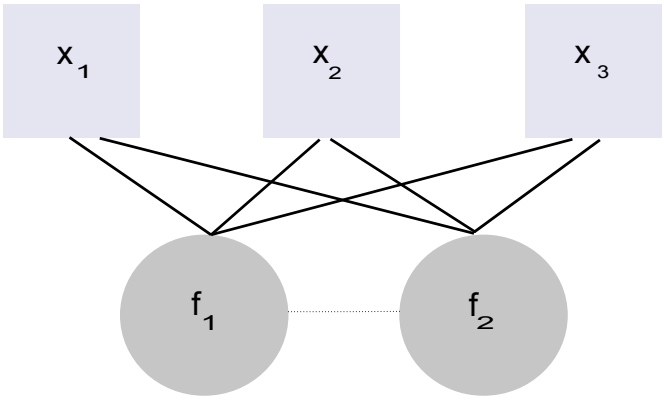


Figure 1.5: The manifest variables x_1, x_2, x_3 are independent of one another conditional on the latent variables f_1, f_2.

variables $\mathbf{f}' = [f_1, \ldots, f_k]$, $k \ll p$, the general latent variable model usually assumes that the relationships observed between the observed variables are independent given the much smaller set of latent variables. This is the key assumption, that the latent variables produce the relationships between the observed variables. Then the behaviour of the observed variables are essentially random, conditioned on the underlying latent variables. Subsequently, all latent variable models assume that the joint probability distribution of $x_1, \ldots, x_p$ conditioned on $\mathbf{f}$, say $\Phi(\mathbf{x}|\mathbf{f})$, is such that $x_1, \ldots, x_p$, given $\mathbf{f}$ are independent. If $\mathbf{x}$ is continuous then Φ is a density function, but if $\mathbf{x}$ is discrete Φ is a set of probabilities. The unconditional density function for $\mathbf{x}$ is obtained by integrating the convolution

$$p(\mathbf{x}) = \int_F \Phi(\mathbf{x}|\mathbf{f})\, h\,(\mathbf{f})\, d\mathbf{f},$$

where $h\,(\mathbf{f})$ is the joint distribution function of $\mathbf{f}$. The conditional probability $\Phi(\mathbf{x}|\mathbf{f})$ is the mapping from the latent variable space to the data space and includes a noise model to account for random error. It is impossible to infer Φ and h uniquely from $p(\mathbf{x})$ without making assumptions. Firstly, as the observed variables are independent, conditioned on the latent variables,

$$\Phi(\mathbf{x}|\mathbf{f}) = \phi_1(x_1|\mathbf{f})\phi_2(x_2|\mathbf{f}) \ldots \phi_p(x_p|\mathbf{f}).$$

Secondly, Φ and h are assumed to be of known form but dependent on a set of unknown parameters which can be identified and estimated. Then inferring Φ and h is a problem of estimating these parameters. The idea is illustrated graphically in Figure 1.5. The observed variables are represented by boxes and the unobserved latent variables by circles, a joining line indicating a dependency between variables. Notice that there can be dependency between the latent variables.

1.3.1 The Factor Analysis Model

The key assumption of the factor analysis model is that given a smaller set of latent variables, the manifest variables $\mathbf{x}$ are essentially independent when conditioned on the latent variables. What this infers is that the observed inter-correlations between the observed variables are explained by the latent variable set except for random error. If they are not, then this indicates that a latent variable is missing from the model or the model is not adequate for the data. This is the basis of what is termed *R-mode* factor analysis. In an R-mode factor analysis the inter-correlations between the observed variables are modelled. A *Q-mode* factor analysis concerns how the objects relate to one another. The R-mode factor model is now discussed. The model is

$$\mathbf{x} = \mathbf{\Gamma f} + \mathbf{e}, \tag{1.5}$$

where $\mathbf{x}$ is a column vector containing the p observed variables, $\mathbf{f} = [f_1, \ldots, f_k]'$ represent the k latent variables ($k \ll p$) or *common factors*, $\mathbf{e} = [e_1, \ldots, e_p]'$ are residual terms and $\mathbf{\Gamma} = [\gamma_{ij}]$ is a $p \times k$ matrix of factor loadings. The model postulates that the observed variables are linear combinations of the latent variables and a residual error. For any given observed variable x_i,

$$x_i = \sum_{j=1}^{k} \gamma_{ij} f_j + e_i. \tag{1.6}$$

For a data sample $\mathbf{X}$, of n observations on p variates, the model becomes,

$$\mathbf{X}_{(n\times p)} = \mathbf{F}_{(n\times k)} \mathbf{\Gamma}'_{(k\times p)} + \mathbf{E}_{(n\times p)}. \tag{1.7}$$

$\mathbf{F}$ is the matrix of factor scores, $\mathbf{\Gamma}$ the matrix of factor loadings and $\mathbf{E}$ is a matrix of residual or error terms.

1.3.2 Variability in the factor model

From equation (1.5) the complete R-mode factor analysis model for the variance and covariance of the manifest variables is given by

$$\mathbf{\Sigma} = \mathbf{\Gamma \Phi \Gamma}' + \mathbf{\Psi}, \tag{1.8}$$

where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{x}$, and $\mathbf{\Phi} = \mathbf{ff}'$ is a $(k \times k)$ matrix, the off diagonal elements contain the correlations between the latent variables. If these are in standardized form, then the main diagonal of $\mathbf{\Phi}$ contains unities. Finally, $\mathbf{\Psi} = [\psi_{ii}]$ is a diagonal matrix containing the variance of the residual errors. Also, if the latent

variables are uncorrelated then $\mathbf{\Phi}$ is an identity matrix and (1.8) implies that the variances of the observed variables may be split into two parts as follows,

$$\sigma_{ii} = \sum_{j=1}^{k} \gamma_{ij}^2 + \psi_{ii}. \tag{1.9}$$

The first term $\sum_{j=1}^{k} \gamma_{ij}^2$ is called the *communality* and is the variance that x_i shares with the other observed variables through the factors. The covariances of the observed variables are given by

$$\sigma_{ij} = \sum_{r=1}^{k} \gamma_{ir}\gamma_{jr} \tag{1.10}$$

and it is only the factors that are involved in these. The matrix of factor loadings $\mathbf{\Gamma}$ is also known as the *factor pattern matrix*. When standardized factors are uncorrelated then, $\mathbf{\Phi} = \mathbf{I}$, and then the pattern matrix gives the covariance of the observed variables with the factors,

$$\text{cov}\,(x_i, f_j) = \gamma_{ij}. \tag{1.11}$$

1.3.3 Indeterminacy

For a single factor $k = 1$ the model described by (1.5) and (1.8) has a unique solution. However for $k > 1$ no unique solution exists. To determine a particular solution the factors have to be referred back to a set of basis vectors. These basis vectors may be orthogonal or oblique. If the chosen reference basis is oblique the factors are correlated and a full description of the solution requires both the factor pattern matrix $\mathbf{\Gamma}$ and the *factor structure* matrix $\mathbf{\Gamma\Phi}$. In which case the coefficients of the pattern matrix are no longer covariances but should be thought of as regression coefficients. The factor model $\mathbf{x} = \mathbf{\Gamma\Gamma'} + \mathbf{\Psi}$, $\mathbf{\Phi} = \mathbf{I}$, has $pk + p = p(k+1)$ parameters but there are $p(p+1)/2$ variances and covariances in $\mathbf{\Sigma}$. The model will only be useful if it has less parameters than there are unique elements of $\mathbf{\Sigma}$, so $p(k+1) \leq p(p+1)/2$, i.e. $k \leq (p-1)/2$. In the case of a single factor ($k = 1$) a unique solution exists up to the sign. The following is an example for $p = 3$

$$\Sigma = \begin{bmatrix} 2 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 10 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In some cases it possible to find solutions where a residual variance is negative and so not statistically valid. This is known as the *Heywood case*. For example, replacing the variance for x_3 with 8 rather than 10 gives the same factor but $\psi_3 = -1$. When the number of factors is greater than one, a unique solution does not exist, since there is

an infinite number of choices for $\mathbf{\Gamma}$. To illustrate, if $\mathbf{f}$ is replaced by $\mathbf{Rf}$ and $\mathbf{\Gamma}$ by $\mathbf{\Gamma R}'$, where $\mathbf{R}$ is an orthogonal rotation matrix. Then

$$\mathbf{x} = \mathbf{\Gamma R}'\mathbf{Rf} + \mathbf{e},$$

and since $\mathbf{R}'\mathbf{R} = \mathbf{I}$ the model is unchanged by these transformations. Such a transformation also leads to the same form for the covariance matrix, since the new factors have a correlation matrix given by

$$\mathbf{Rff}'\mathbf{R}' = \mathbf{R\Phi R}'$$

so that the new factors $\mathbf{Rf}$ and new loadings $\mathbf{\Gamma R}$ imply that

$$\mathbf{\Sigma} = (\mathbf{\Gamma R}')(\mathbf{R\Phi R}')(\mathbf{R\Gamma}') + \mathbf{\Psi},$$

$$= \mathbf{\Gamma\Phi\Gamma}' + \mathbf{\Psi}.$$

Consequently, the ability of the new factors to account for the variances and covariances of the observed variables is exactly equivalent to the original factors. Such a transformation corresponds to a rotation of the factors.

1.3.4 Confirmatory Factor Analysis

In some situations an investigator may wish to fix certain parameters in $\mathbf{\Gamma}$ and $\mathbf{\Psi}$. This is then termed a *confirmatory factor analysis* and this may lead to a unique solution for the free parameters as a rotation would destroy the pattern of the fixed parameters. If the number of fixed parameters are denoted respectively by $n_{\mathbf{\Gamma}}$ and $n_{\mathbf{\Psi}}$ then a necessary but not sufficient condition for a unique solution is

$$n_{\mathbf{\Gamma}} + n_{\mathbf{\Psi}} \geq k^2.$$

However, in general it is difficult to give sufficient conditions for uniqueness, since the position of the fixed parameters is also important.

1.3.5 Estimating the Model Parameters

To estimate the parameters, a discrepancy function between the parametrized model covariance matrix $\mathbf{\Sigma}(\mathbf{\Gamma}, \mathbf{\Psi})$ and the unbiased sample covariance matrix $\mathbf{S}$ is minimized. Commonly, *maximum likelihood* is used, but other discrepancy functions are possible, for example, ordinary least squares and generalized least squares. See Everitt for details. The aim is to estimate $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ so that

$$\mathbf{S} = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}' + \hat{\mathbf{\Psi}},$$

where the hat symbol above a parameter, matrix or vector indicates that it is estimated. Here the factors are orthogonal so that $\mathbf{\Phi} = \mathbf{I}$. If $\mathbf{x}$, $\mathbf{f}$ and $\mathbf{e}$ have multivariate normal distributions then maximizing the log-likelihood is equivalent to minimizing the discrepancy function

$$F\left(\mathbf{S}, \mathbf{\Sigma}\left(\mathbf{\Gamma}, \mathbf{\Psi}\right)\right) = \log_e |\mathbf{\Sigma}| + \text{trace}\mathbf{S}\mathbf{\Sigma}^{-1} - \log_e |\mathbf{S}| - p$$

with $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}$. This can be minimized using an iterative procedure suggested by Jöreskog (see Mardia et al., 1979, for the detail).

1.3.6 Principal Component Factor Analysis

A PCA is primarily a dimensionality reduction technique. However, if the principal components $\mathbf{y} = \mathbf{A}\mathbf{x}$ are now considered to be factors then the principal components can be reformulated as a factor analysis model. The model becomes, by multiplying by the inverse of $\mathbf{A}$

$$\mathbf{x} = \mathbf{A}'\mathbf{y}. \tag{1.12}$$

If the first k components describe the variation in the data that captures the relationships between the observed variables, then the remaining components represent the residual variation or random error, $\mathbf{e}$ and the model can be written

$$\mathbf{x} = \mathbf{A}'_k\mathbf{y}_k + \mathbf{e}.$$

These errors are taken to be uncorrelated. In terms of the covariance of $\mathbf{x}$, this gives

$$\begin{aligned}
\mathbf{\Sigma} &= \text{var}\left(\mathbf{A}'_k\mathbf{y}_k + \mathbf{e}\right) \\
&= \text{var}\left(\mathbf{A}'_k\mathbf{y}_k\right) + \text{var}\left(\mathbf{e}\right) + 2\text{cov}\left(\mathbf{A}'_k\mathbf{y}_k, \mathbf{e}\right) \\
&= \text{var}\left(\mathbf{A}'_k\mathbf{y}_k\right) + \mathbf{\Psi},
\end{aligned}$$

assuming $\text{cov}\left(\mathbf{A}'_k\mathbf{y}_k, \mathbf{e}\right) = \mathbf{0}$. Then

$$\mathbf{\Sigma} = \mathbf{\Gamma}'\mathbf{\Phi}\mathbf{\Gamma} + \mathbf{\Psi},$$

where $\mathbf{\Gamma} = \mathbf{A}_k$, $\mathbf{\Phi} = \mathbf{\Delta}_k$ and $\mathbf{\Psi}$ is a diagonal matrix of errors.

The communality $\hat{h}_i$ of the ith observed variable is defined across a subset of k factors as

$$\hat{h}_i = \sum_{j=1}^{k} a_{ji}^2,$$

where a_{ji} is the ith loading of the jth principal component. The correlation of the observed variables with each of the factors is important after a PCA on the covariance matrix, as the magnitude of the loadings, will not in general, represent these correlations, particularly if the variables are measured on different scales. For a PCA on

a correlation matrix the variable loadings on the factors reflect its correlation with a given factor. The covariance of an observed variable x_i with a factor y_j is given as

$$\begin{aligned} \text{cov}\,(y_j, x_i) &= \text{cov}\left(y_j, \sum_{k=1}^{p} a_{ik} y_k\right) \\ &= \sum_{k=1}^{p} a_{ik} \text{cov}\,(y_j, y_k) \\ &= a_{ij} \text{var}\,(y_j) \\ &= a_{ij} \lambda_j \end{aligned}$$

and the correlation is given as

$$\begin{aligned} \text{corr}\,(y_j, x_i) &= \frac{a_{ij}\lambda_j}{\lambda_j^{\frac{1}{2}}} \\ &= a_{ij}\lambda_j^{\frac{1}{2}}. \end{aligned}$$

So in general

$$\text{corr}\,(\mathbf{y}, \mathbf{x}) = \mathbf{A}\boldsymbol{\Lambda}^{\frac{1}{2}}. \tag{1.13}$$

1.3.7 Factor Rotation

As mentioned in Section 1.3.3, factors are not unique and any rotation of a factor subset is also a solution. Rotation of a subset of the principal components to a simpler structure will conserve the total variance explained but it becomes more spread between components with either a loss of ortho-normality or the introduction of correlation. What is the best way to choose a solution? One approach is to make factors align more with the original variables, i.e. making a few of the coefficients within factors as large as possible in magnitude, and the rest small. This can be achieved by applying additional constraints on the optimization.

Rotation methods have attracted much criticism because the choice of rotation technique can often affect the final interpretation of the analysis. However, according to Everitt these criticisms overlook two important points. Firstly, although the axes may be rotated about their origin, or become oblique, the distribution of points will remain invariant. If the loadings are found to be in groups or concentrated in one or two parts of the space, then it is reasonable to choose new axes in a way which will allow the positions of these loadings to be described as simply as possible, that is using as few parameters as possible. Secondly, the rotation methods are of primary relevance when the investigation is exploratory in nature; in these situations the hope is that the use of factor analysis methods will allow the experimenter to formulate hypotheses which can then be submitted to testing on further data by a confirmatory analysis.

The Orthogonal Rotation

An orthogonal rotation preserves the orthogonality between all the factors but will induce correlation, unless the factors are standardized by scaling with their standard deviations, $\mathbf{\Delta}^{-\frac{1}{2}}$. A *varimax* rotation (Kaiser, 1958) is a commonly applied orthogonal rotation. The varimax method searches for a linear combination of a subset of the original factors, such that the sum of the variances of the squared loadings is maximized,

$$V = \sum_{j=1}^{k} \left(\sum_{i=1}^{p} a_{ij}^2 - \frac{1}{p} \sum_{i=1}^{p} a_{ij}^2 \right)^2 ,$$

with a_{ij}^2 being the squared loading of the ith variable on the jth factor, and k the number of factors which are rotated.

There are many orthogonal rotations available, for example others are *quartimax* and *orthomax* rotations. All try to align the new orthogonal axes with the variables. Orthogonal rotations are commonly used, and will locate orthogonal variable clusters. However, axes that align better with natural variable clusters may better fulfill Thurstone's criteria (see Section 2.1). Oblique rotations relax the condition that factors must be orthogonal and allow the new axes to take any position in the factor space.

The Oblique Rotation

Methods which relax independence and allow correlated factors are termed *oblique*, for example, the *Oblimin* (Harman, 1976) and *Promax* (Hendrickson and White, 1964). Promax is a computationally efficient method which attempts to further polarize the loadings from an orthogonal position. Harris and Kaiser (1964) give a set of methods called *Orthoblique*, which are invariant under rescaling of the axes. Recently, Jennrich (2002) proposed a gradient projection algorithm as a general method for oblique rotation. Many oblique methods were developed in the psychometric literature, however, according to Basilevsky (1994), not a great deal of statistical and numerical work has been done. Authors such as Fabrigar et al. (1999) and Costello and Osborne (2005), recommend oblique rotations as best practice in exploratory factor analysis. These authors make a strong argument in favour of oblique rotations rather than orthogonal solutions. They note that dimensions of interest are not often dimensions that would be orthogonal. If the latent variables are, in fact, correlated, then an oblique rotation will produce a better estimate of the true factors and a better simple structure than will an orthogonal rotation. If the oblique rotation indicates that the factors have close to zero correlations between one another, then the analyst can go ahead and conduct an orthogonal rotation (which should then give about the same solution as the oblique rotation).

This thesis explores factor rotations that preserve orthogonality while maximizing the correlation between the factors. The R-mode factor analysis model can be formulated in two ways; as a *true factor analysis* in which factors account for the maximum inter-correlations of the observed variables and *principal component factor analysis* where factors account for maximum variance. The next section highlights some key ideas around the rotation of principal component factors. This links into the work in later chapters.

1.3.8 Principal Components and Factor Rotation

Orthogonal Rotation of Principal Components

Jolliffe (1995) discusses the effects of the orthogonal rotation of principal components and shows why it is not possible to preserve rotated components which are pairwise uncorrelated and/or whose loadings are orthogonal. Consider the mean centred or standardized data sample $\mathbf{X}$, then its principal components are given by

$$\mathbf{Y} = \mathbf{X}\,\mathbf{U},$$

using the spectral decomposition of the covariance matrix of $\mathbf{X}$ (Section 1.1.1). Taking the first k components and treating the remaining components as residual error, $\mathbf{e}$, a PC factor model can be written

$$\mathbf{X} = \mathbf{Y}_k\mathbf{U}_k' + \mathbf{e}, \tag{1.14}$$

and the covariance matrix of $\mathbf{X}$ is modelled as

$$\mathbf{\Sigma} = \mathbf{U}_k\mathbf{Y}_k'\mathbf{Y}_k\mathbf{U}_k' + \mathbf{\Psi},$$

$\mathbf{\Psi}$ denoting a diagonal matrix of residual variance. As $\mathbf{Y}_k$ are principal components, $\mathbf{Y}_k'\mathbf{Y}_k = \mathbf{\Delta}_k$, which is a diagonal matrix of the first k eigenvalues of $\mathbf{\Sigma}$ in descending order of magnitude. Then,

$$\mathbf{\Sigma} = \mathbf{U}_k\mathbf{\Delta}_k\mathbf{U}_k' + \mathbf{\Psi}.$$

Notice that the factors, $\mathbf{Y}_k$ and the principal component loading vectors are uncorrelated as $\mathbf{U}_k'\mathbf{U}_k = \mathbf{I}$ and $\mathbf{Y}_k'\mathbf{Y}_k = \mathbf{\Delta}_k$. As mentioned earlier the model is invariant to an orthogonal rotation of the principal axes. Let $\mathbf{R}$ be an orthogonal rotation, then $\mathbf{R}'\mathbf{R} = \mathbf{R}\mathbf{R}' = \mathbf{I}$ and the model becomes

$$\mathbf{X} = (\mathbf{Y}_k\mathbf{R})(\mathbf{U}_k\mathbf{R})' + \mathbf{e}$$

which is equivalent to (1.14). However, the factors will no longer remain uncorrelated, as

$$(\mathbf{Y}_k\mathbf{R})'(\mathbf{Y}_k\mathbf{R}) = \mathbf{R}'\mathbf{Y}_k'\mathbf{Y}_k\mathbf{R} = \mathbf{R}'\mathbf{\Delta}\mathbf{R}.$$

The factor loadings will remain orthogonal,

$$(\mathbf{U}_k\mathbf{R})'(\mathbf{U}_k\mathbf{R}) = \mathbf{R}'\mathbf{U}_k'\mathbf{U}_k\mathbf{R} = \mathbf{I}.$$

In practice the factors are usually standardized, which causes the factors to remain uncorrelated after an orthogonal rotation (the loadings become non-orthogonal). This practice is criticized in the literature as standardization effectively stretches the scores to sit on a hypersphere so that any position of the axes will not induce correlation.

To standardize the factors, let $\mathbf{Z} = \mathbf{Y}\boldsymbol{\Delta}^{-\frac{1}{2}}$, and the factor model becomes,

$$\mathbf{X} = \mathbf{Z}_k\boldsymbol{\Delta}_k^{\frac{1}{2}}\mathbf{U}_k' + \mathbf{e}$$

and

$$\boldsymbol{\Sigma} = \mathbf{U}_k\boldsymbol{\Delta}_k^{\frac{1}{2}}\mathbf{Z}_k'\mathbf{Z}_k\boldsymbol{\Delta}_k^{\frac{1}{2}}\mathbf{U}_k' + \boldsymbol{\Psi}.$$

Now, $\mathbf{Z}_k'\mathbf{Z}_k = \mathbf{I}_k$ and the factor loadings are $\mathbf{U}_k\boldsymbol{\Delta}_k^{\frac{1}{2}}$. So after an orthogonal rotation of the axes, the factors remain uncorrelated as,

$$(\mathbf{Z}_k\mathbf{R})'(\mathbf{Z}_k\mathbf{R}) = \mathbf{I}_k$$

but the loadings are no longer uncorrelated, as

$$(\mathbf{U}_k\boldsymbol{\Delta}_k^{\frac{1}{2}}\mathbf{R})'(\mathbf{U}_k\boldsymbol{\Delta}_k^{\frac{1}{2}}\mathbf{R}) = \mathbf{R}'\boldsymbol{\Delta}_k\mathbf{R}.$$

In the literature, if factors are highly correlated this is taken as meaning the factors should really be one single factor. Oblique rotations, will better align with natural variable clusters and for this reason are recommended. However, in certain circumstances it may be useful to obtain orthogonal factors which are highly correlated. For example, when groups of correlated components differentiate to describe a latent trait in different ways. Or given a configuration of points from an analysis, for example an MDS, and the axes are arbitrary, it would be useful to rotate the configuration in such a way that the correlation or covariance between the latent variables is maximized, but keeping the axes orthogonal. In this way the latent variables could be displayed on a parallel coordinate plot. The axes remain independent, but the plot becomes easier to interpret as groups differentiate and the number of cross-overs on the plot are minimized. These applications are investigated in Chapter 3.

Oblique Rotation of Principal Components

As mentioned briefly in the last section, an oblique rotation will better align the factors with natural variable clusters. An application of this is to identify variable clusters and is discussed in Section 2.9, which looks at a method to obtain a simple interpretation of a large data set. A brief overview of oblique rotations is given here for reference.

Firstly, an oblique rotation relaxes the requirement that the axes are orthogonal, and so finding oblique axes is more akin to regression. Basilevsky has the detail. If both the variables and factors are standardized to unit length, then, if $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k\}$ represents the oblique basis then,

$$\hat{\mathbf{\Sigma}} = \mathbf{B}\mathbf{\Phi}\mathbf{B}' + \mathbf{\Psi}$$

and

$$\mathbf{\Phi} = \mathbf{G}'\mathbf{G}$$

which is the correlation matrix of the oblique axes. $\mathbf{B}$ is described in terms of an ordinary least squares projection of the data $\mathbf{X}$,

$$\mathbf{B}' = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{X} \tag{1.15}$$

and so the estimate of $\mathbf{X}$ is

$$\hat{\mathbf{X}} = \mathbf{G}\mathbf{B}' = \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{X} = \mathbf{P}_{\mathbf{G}}\mathbf{X}$$

where $\mathbf{P}_{\mathbf{G}}$ is an idempotent, symmetric projection matrix. From (1.15),

$$\mathbf{\Phi}\mathbf{B}' = \mathbf{G}'\mathbf{X}.$$

$\mathbf{G}'\mathbf{X}$ is the correlation matrix of the variables and the oblique components, called the matrix of correlation loading coefficients. $\mathbf{B}$ are the regression coefficients and represent the coordinates of $\mathbf{X}$ with respect to the oblique components $\mathbf{G}$. $\mathbf{G}$ is not unique and to define the oblique basis a further constraint is required. Criterion such as *oblimin* provide this, and in a similar way to the varimax criterion guide the axes position to align with the variables.

Chapter 3 explores the case where axes can be found which remain orthogonal but the induced correlation between factors may provide groups of axes which although correlated, describe different aspects of a latent trait.

1.3.9 The Sensory Panel Example Re-visited

In Section 1.1.3, a subset of five principal components were selected to represent the data. Unfortunately, many small loadings were present on the loading vectors, and individual vectors could not be easily labelled. Table 1.7 shows the five loading vectors after a varimax rotation. It is now easier to label the individual components, Table 1.5. For instance RL1 contrasts drying with tactile, RL2, is an average of applicator use with tactile, RL3, drying with efficacy, RL4, drying with efficacy and tactile and RL5, fragrance. However, there are still many small non-informative loadings and subjective decisions have been made in labelling the groups and rotated components. None of the

rotated component scores correlate with overall opinion as highly as did the first PC, Table 1.6, but a regression model using the new factors will be more interpretable. As mentioned previously, Chapter 2 reviews approaches to deal with the subjective choice of loadings, and a new algorithm is proposed.

Group	Description	RL1	RL2	RL3	RL4	RL5
1	Use of product applicator		+			
2	Drying and deposits	-		-	-	
3	Fragrance					+
4	Odour and wetness efficacy			+	+	
5	Tactile properties	+	+		+	

Table 1.5: A subjective interpretation of the loadings on the rotated principal components

Rotated Factor	Correlation with overall opinion
Factor1	0.46
Factor2	0.27
Factor3	0.39
Factor4	0.08
Factor5	0.38

Table 1.6: The correlation of the rotated factor scores with overall opinion for the deodorant data

1.3.10 Other Latent Variable Models

In recent years a number of unsupervised learning algorithms have emerged from the machine learning literature, with an emphasis on their probabilistic framework. Density networks were proposed by MacKay and Gibbs (1999). These do not impose any fixed probability distribution on the latent variable model, but instead use Monte Carlo simulation within a Bayesian framework to learn the intrinsic data manifold. However, because of the need to take Monte Carlo samples in the latent space, these grow exponentially with the dimension of the latent space. Svensen (1998) proposed the generative topographic mapping (GTM), following the LVM framework. It uses a non-linear mapping from the latent space to the data space in the form of generalized linear model and uses a mixture of Gaussian distributions to model the induced manifold in data space. It represents a probabilistic alternative to the self organizing map (SOM). A SOM is an artificial neural network which is trained to produce a low dimensional topologically ordered map Kohonen (1995). Various other forms of the LVM model exist including independent component analysis, probabilistic principal component analysis, independent factor analysis. These all fit the general latent variable framework mentioned but take different forms for the prior, noise and mapping.

Code	Description	Group	RL1	RL2	RL3	RL4	RL5
A1	rollball glided over skin	1	0.11	0.82	0.13	0.03	0.07
A11	ball rolled freely in pack		0.11	0.80	0.06	0.02	-0.03
A12	ball did not dry out		0.04	0.65	0.11	-0.01	-0.04
A13	pack did not become messy		0.08	0.54	0.02	0.37	-0.01
A14	product did not leak out		0.09	0.46	0.00	0.36	-0.08
A16	easy of application		0.14	0.73	0.15	0.17	0.09
A19	how smooth whilst applying		0.18	0.61	0.15	0.18	0.15
A39	overall opinion packaging		0.07	0.42	0.10	0.23	0.15
A17	how product dosed from pack	2	-0.46	0.27	-0.04	-0.04	0.10
A43	felt wet during application		-0.82	-0.12	-0.10	-0.03	-0.03
A44	felt sticky whilst drying		-0.77	-0.22	-0.10	-0.13	-0.02
A45	left visible deposits		-0.20	-0.23	-0.24	-0.62	-0.14
A46	cold on application		-0.35	-0.24	-0.14	-0.32	-0.12
A47	marked clothes		-0.23	-0.14	-0.19	-0.69	0.00
A48	waited longer than usual- drying		-0.81	-0.13	-0.09	-0.14	-0.08
A49	felt greasy		-0.59	-0.30	-0.07	-0.40	-0.04
A33	overall opinion fragrance	3	0.16	0.17	0.03	0.08	0.80
A34	strength fragrance-immediately		-0.09	-0.08	0.05	-0.04	0.69
A35	strength fragrance- end of day		-0.01	0.04	0.31	-0.05	0.70
A6	had a pleasant fragrance		0.10	0.18	-0.02	0.14	0.79
A7	fragrance lasted long enough for me		0.12	0.07	0.36	0.05	0.77
A10	kept me fresh all day	4	0.23	0.19	0.80	0.12	0.23
A28	overall opinion - effective		0.41	0.33	0.60	0.18	0.28
A29	notice any perspiration		0.12	0.03	0.72	0.16	0.00
A30	overall how effective keeping you dry		0.22	0.14	0.80	0.23	0.07
A31	notice any odour		-0.02	0.09	0.63	0.05	0.11
A32	how effective keeping free from odour		0.08	0.19	0.75	0.17	0.20
A36	notice visible deposits - skin		0.10	0.00	0.16	0.73	-0.02
A37	notice deposits on clothes		0.08	-0.06	0.15	0.69	-0.01
A38	how easy to wash off skin		0.16	0.09	-0.12	0.28	0.08
A40	any irritation		0.18	0.12	0.02	0.02	0.09
A41	any trapping of underarm hair		0.03	0.21	0.03	0.27	0.02
A42	how often applied rollon		-0.02	-0.03	0.09	-0.04	-0.04
A8	gave me daylong protection - BO		0.16	0.19	0.80	0.10	0.20
A9	gave me daylong protection- wetness		0.24	0.14	0.83	0.15	0.07
A15	easy to apply the right amount	5	0.36	0.50	0.01	0.08	0.05
A18	ease of applying right amount		0.38	0.41	0.14	0.06	0.10
A2	felt fresh whilst applying		0.31	0.48	0.14	0.09	0.36
A20	how sticky whilst applying		0.72	0.24	0.06	0.22	0.06
A21	how greasy whilst applying		0.46	0.37	0.07	0.33	-0.03
A22	how wet whilst applying		0.79	0.03	0.10	0.09	0.03
A23	how cold whilst applying		0.40	0.14	0.11	0.33	0.06
A24	how sticky immediately after application		0.73	0.21	0.05	0.20	0.04
A25	speed of drying		0.81	0.11	0.15	0.04	0.00
A26	how sticky whilst wearing		0.46	0.24	0.16	0.40	0.13
A27	how greasy whilst wearing		0.32	0.36	0.17	0.43	0.05
A3	felt smooth whilst applying		0.35	0.60	0.10	0.05	0.23
A4	dried quickly		0.85	0.10	0.11	0.14	0.05
A5	left underarm soft and smooth		0.47	0.29	0.19	0.18	0.30

Table 1.7: The rotated loading vectors for the deodorant data example, section 1.1.3

1.4 The Statistical Analysis of Shape

In Chapter 4, the utility of a questionnaire where the response to a question is recorded in a two dimensional space is explored. In this case respondents do not explicitly score each object on a linear scale, but rather perform a multiple comparison in the two dimensional space. The intention is to illicit information that is not consciously expressed, so called tacit information. Such a response can be thought of as a shape configuration.

Kendall (1984) pioneered statistical shape analysis using *point configurations*. In essence, translation, scale and rotation (the Euclidean similarity transformations) are nuisance parameters that need to be removed. The analysis of shape can be performed in a coordinate system or using a non-coordinate approach where, the distance between points, termed *landmarks*, represent the configuration. In which case the shape configuration becomes invariant to translation, rotation and reflection. To use a coordinate system, configurations must first be *registered* into that system, referred to as a shape space. This approach follows that detailed in Dryden and Mardia (1998). The second option is to use a representation of shape that is coordinate free. The coordinate free approach is detailed by Lele and Richtsmeier (2000). A coordinate free system, based on a measure of distance between landmarks has certain advantages over using a coordinate system.

1.4.1 Shape Coordinate Systems

Registration is the process of removing nuisance parameters by translating, scaling and rotating shapes into a common shape coordinate system. Many coordinate systems have been proposed. See Bookstein (1984, 1986), Kendall (1984), Watson (1986). Kendall proposed a spherical coordinate system which results in a *Non-Euclidean shape space*. In general there are k labelled points in m real dimensions. Let the $k \times m$ matrix X denote a landmark configuration. If G is defined as a group of operations acting on X, called a registration group, then G may be one of the following,

- Euclidean similarity group (translation, scale, rotation, reflection)
- Isometry group (translation and rotation)
- Affine group (translation, rotation and shears) .

The first of these is the required group of transformations, as the removal of shear would lead to a loss of information. Unfortunately, the registration process leaves

artefacts. For example, Bookstein coordinates fix or register two landmarks of a given configuration leaving the remaining $k-2$ landmarks. A consequence of this approach is that the application of a Euclidean distance metric results in inconsistent shape similarities. In fact a non Euclidean metric is required (Bookstein, 1986).

The process can be illustrated by Kendall's shape space which involves the following steps. Location is removed by centring the landmark configurations.

$$\mathbf{X}_c = \mathbf{C}\mathbf{X},$$

$\mathbf{C}$ is a centring matrix $\mathbf{C} = \mathbf{I}_k - \frac{1}{k}\underline{\mathbf{1}}_k\underline{\mathbf{1}}_k'$. Size is removed by rescaling with the centroid size,

$$\mathbf{Z} = \frac{\mathbf{X}_c}{S(\mathbf{X})} = \frac{\mathbf{C}\mathbf{X}}{S(\mathbf{X})}.$$

This is called *Pre-shape*. The centroid size is the square root of the mean of the Euclidean squared distances from the centroid of the shape to the landmarks and is given by $S(X) = \|\mathbf{C}\mathbf{X}\|$. Finally the pre-shapes are rotated to get shape

$$[\mathbf{X}] = \{\Gamma : Z\Gamma \in SO\{m\}\}$$

$SO(m)$ is the special orthogonal group of matrices, and $[\mathbf{X}]$ denotes the shape of $\mathbf{X}$.

In general, translation may be removed by pre-multiplying the configuration with a suitable matrix.

$$\mathbf{M}(\mathbf{X} + \bar{\mathbf{t}}) = \mathbf{M}\mathbf{X} \quad \forall \quad \bar{\mathbf{t}},$$

where $\bar{\mathbf{t}}$ is any translation $\in \mathbb{R}^m$. One option, mentioned previously, is the centring matrix.

Shape Distances

In order to do geometry in the shape space a defined distance metric is required. Consider two shapes $[\mathbf{X}_1]$ and $[\mathbf{X}_2]$ with pre-shapes $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The full Procrustes distance between them is defined as

$$d_F([\mathbf{X}_1], [\mathbf{X}_2]) = \min_{r>0,\Gamma} \|\mathbf{Z}_2 - r\mathbf{Z}_1\Gamma\|$$

this represents the shortest distance between two points on the shape sphere, which is a great circle. This is

$$d_F([\mathbf{X}_1], [\mathbf{X}_2]) = \left\{1 - \left(\sum_{i=1}^{m} \lambda_i\right)^2\right\}^{\frac{1}{2}}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{m-1} \geq \lambda_m$ are the square roots of the eigenvalues of $\mathbf{Z}_1'\mathbf{Z}_2\mathbf{Z}_2'\mathbf{Z}_1$. The minimizing rotation is given by

$$\hat{\mathbf{\Gamma}} = \mathbf{U}\mathbf{V}',$$

where $\mathbf{U}, \mathbf{V} \in \mathrm{SO}\{m\}$ and $\mathbf{Z}_2'\mathbf{Z}_1 = \mathbf{V}\mathbf{\Delta}\mathbf{U}'$ with $\mathbf{\Delta} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$. The minimizing scale is $\hat{r} = \sum_{i=1}^{m} \lambda_i$, (see, Kendall, 1984, Dryden and Mardia, 1998).

Table 1.8 summarises Procrustes distance in shape space.

- Partial Procrustes distance, where scale is not removed

$$d_P(\mathbf{X}_1, \mathbf{X}_2) = \min_{\Gamma \in SO\{m\}} \|\mathbf{Z}_2 - \mathbf{Z}_1\Gamma\|$$

- Full Procrustes distance

$$d_F(\mathbf{X}_1, \mathbf{X}_2) = \min_{r>0,\Gamma} \|\mathbf{Z}_2 - r\mathbf{Z}_1\Gamma\| = \sin\rho(\mathbf{X}_1, \mathbf{X}_2)$$

- Riemannian distance

$$\rho(\mathbf{X}_1, \mathbf{X}_2) = 2\arcsin(d_{12}/2)$$

  d_{12} is the Euclidean distance between $\mathbf{X}_1$ and $\mathbf{X}_2$

Distance	Notation	Formula	Range
Full Procrustes distance	d_F	$\left\{1 - (\sum_{i=1}^{m} \lambda_i)^2\right\}^{\frac{1}{2}}$	$0 \leq d_F \leq 1$
Partial Procrustes distance	d_P	$\sqrt{2}(1 - \sum_{i=1}^{m} \lambda_i)^{\frac{1}{2}}$	$0 \leq d_P \leq \sqrt{2}$
Riemannian distance	ρ	$\arccos(\sum_{i=1}^{m} \lambda_i)$	$0 \leq \rho \leq \frac{\pi}{2}$

Table 1.8: Procrustes distances in shape space, taken from Dryden and Mardia (1998)

Tangent Space Coordinates and PCA

A more advanced coordinate system to analyse shape is the tangent space. For a definition of tangent space please refer to O'Neill (1997). This can be thought as a linearised version of the shape space. The tangent space coordinates are obtained by a generalized Procrustes alignment, followed by a projection of the full Procrustes coordinates into the tangent space about a pole, this is chosen to be the full Procrustes mean, for details see Dryden and Mardia. Multivariate techniques in tangent space involving distances to the pole are equivalent to non-Euclidean shape methods requiring Procrustes distances, provided that the data is not too highly dispersed.

For the analysis and interpretation of multivariate observations, a standard method which has been used in the tangent space is PCA, for example Cootes et al. (1992).

1.4.2 Procrustes methods

Procrustes matching has two contexts, matching matrices and matching shape data. Procrustes techniques were pioneered, initially in the field of psychology (Mosier, 1939). Details can be found in Cox and Cox (2000), Gower and Hand (1996). *Ordinary Procrustes analysis* is the process of matching one matrix to another. The more general process of matching many configurations is known as *general Procrustes analysis* which is an iterative method pioneered by Gower (1975) and Berge (1977), but adapted explicitly for shapes by Goodall (1991).

1.4.3 An Invariant Approach for the Analysis of Shape

The use of landmark configurations within a coordinate system to describe and analyse shapes statistically, suffers from many difficulties. For instance, non-Euclidean distance metrics and the effects of constrained Euclidean nuisance parameters. The use of a coordinate free approach overcomes these problems. The matrix of Euclidean distances between landmarks (EDM) is used to define shape. For a landmark configuration the metric distance between labelled points after a suitable normalization is used to define its shape. Lele and Richtsmeier (2000) call the matrix of distances between landmarks a *form matrix*. As an example consider a simple equilateral triangle $\mathbf{X}$ with the following landmark configuration,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 1.5 & \frac{\sqrt{3}}{2} \end{pmatrix}.$$

Then the form matrix for $\mathbf{X}$ can simply be written as the matrix of Euclidean distances between landmarks,

$$FM(\mathbf{X}) = \begin{pmatrix} 0 & d_{12} & d_{13} \\ d_{21} & 0 & d_{23} \\ d_{31} & d_{32} & 0 \end{pmatrix}.$$

As $d_{ij} = d_{ji}$ the unique elements can be taken to give the simplified vector,

$$FM(\mathbf{X}) = \begin{pmatrix} d_{12} \\ d_{13} \\ d_{23} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

The mean form of a sample is derived from the mean of the corresponding inter-landmark distances. The nuisance parameters of translation, rotation and scale are irrelevant. Unfortunately, estimating shape variability still suffers from these nuisance

parameters, but estimators of population moments are statistically consistent. That is, as the sample size increases the probability that the estimate approaches the true population moment will increase.

Comparing Form and Shape

As mentioned, the form matrix is purely the inter-landmark Euclidean distances and so includes size. To get to shape with this approach, the form matrix is scaled using the geometric mean of the distances,

$$S(\mathbf{X}) = \left\{ \prod FM_{ij}(\mathbf{X}) \right\}^{\frac{1}{L}}.$$

L is the dimension of the form space, i.e. the number of inter-landmark distances,

$$L = \frac{k(k-1)}{2}$$

and FM_i is each inter-landmark distance. So, the shape matrix (SM) is

$$SM(\mathbf{X}) = \frac{FM(\mathbf{X})}{S(\mathbf{X})},$$

and the difference in size between two forms is,

$$S_{diff} = \frac{S(\mathbf{X})}{S(\mathbf{Y})}.$$

As the form space is still Euclidean, differences in shapes are easily quantified,

$$SDM_i(\mathbf{X}, \mathbf{Y}) = SM_i(\mathbf{X}) - SM_i(\mathbf{Y}),$$

where SDM is the shape difference matrix and SDM_i is shorthand for subtracting each of the corresponding inter-landmark distance between $SM(\mathbf{X})$ and $SM(\mathbf{Y})$. If the mean shapes of two samples are denoted by $SM(\bar{\mathbf{X}})$ and $SM(\bar{\mathbf{Y}})$, then to calculate the difference between mean shapes,

$$SDM_i(\bar{\mathbf{X}}, \bar{\mathbf{Y}}) = SM_i(\bar{\mathbf{X}}) - SM_i(\bar{\mathbf{Y}}).$$

The Gaussian Noise Model

Quantifying the variability for landmark configurations is more convoluted. The variability is still dependent on the Euclidean transformations required to match the con-

figurations. There are two kinds of variability to consider. The variability between dimensions and the variability between points. A convenient way to represent this structure is the *Kronecker product*. $\boldsymbol{\Sigma}_k \otimes \boldsymbol{\Sigma}_m$ which splits the variance into components for the k points and m dimensions.

For a Gaussian noise model a configuration has the following distribution,

$$\mathbf{X} \sim N(\bar{\mathbf{X}}, \boldsymbol{\Sigma}_K \otimes \boldsymbol{\Sigma}_m).$$

If translation, t and rotation, $\mathbf{R}$, are applied to the configuration then,

$$\mathbf{X}\mathbf{R} + \underline{1}\mathbf{t} \sim N(\bar{\mathbf{X}}\mathbf{R} + \underline{1}\mathbf{t}, \boldsymbol{\Sigma}_k \otimes \mathbf{R}'\boldsymbol{\Sigma}_m\mathbf{R}),$$

which indicates that translation and rotation of a configuration affects both its mean and its covariance matrix.

As mentioned, the estimation of parameters for a Gaussian model leads to consistent estimators in the case of an EDM representation, Lele (1993). However, there are identifiability issues. For the EDM approach, the following results hold.

Result 1: The mean landmark configuration matrix $\bar{\mathbf{X}}$ cannot be estimated. However, $FM(\bar{\mathbf{X}})$ can be estimated.

Result 2: Neither $\boldsymbol{\Sigma}_k$ nor $\boldsymbol{\Sigma}_m$ can be estimated. Only a singular version of $\boldsymbol{\Sigma}_k$ denoted as $\boldsymbol{\Sigma}_k^*$ and only the eigenvalues of $\boldsymbol{\Sigma}_m$ can be estimated.

$\boldsymbol{\Sigma}_k^*$ and $\boldsymbol{\Sigma}_m$ are termed the *perturbation pattern*. The estimation of the exact quantities for the local landmark variation is impossible due to the nuisance parameters.

To summarize, the use of the EDM form representation gives consistent parameter estimates under a Gaussian noise model. The form matrix is invariant to reflection and the form matrix approach partitions configurations into equivalence classes, called orbits. This is unambiguous, not suffering from the effects of the nuisance parameters.

Methods to explore the form difference matrix

Often an investigator is interested in local differences between particular landmarks. Lele and Richtsmeier (1992) and Lele and Cole (1996) present some tools to investigate these differences. A couple of approaches to detect influential landmarks are outlined.

1. Landmark deletion approach
   Differences between forms can be quantified in terms of the relative form differ-

ence. In which case, the form difference matrix of two forms $FDM(\mathbf{B}, \mathbf{A})$ consists of the ratios of the elements of the two forms. Consequently, an element close to one indicates that the landmark distances are similar between the two forms.

$$FDM(\mathbf{B}, \mathbf{A}) = \frac{FM_i\mathbf{B}}{FM_i\mathbf{A}}$$

Elements of $FM(\mathbf{A})$ and $FM(\mathbf{B})$ are inter-landmark distances. The following test statistic is used

$$T = \frac{\max_i FDM_i(\mathbf{B}, \mathbf{A})}{\min_i FDM_i(\mathbf{B}, \mathbf{A})}$$

to detect influential landmarks.

(a) Calculate T for all landmarks

(b) For $i = 1, \ldots, k$ delete the ith landmark and recalculate T to give T_{-1}, T_{-2}, $\ldots T_{-k}$
If T exhibits a large drop then this is indicative of influential landmarks
This procedure can be also be applied to groups of landmarks.

2. Graphical tool for detecting influential landmarks
The idea is to plot for each landmark the elements of the $FDM(\mathbf{B}, \mathbf{A})$ that include that landmark. This corresponds to rows of the $FDM(\mathbf{B}, \mathbf{A})$. Influential landmarks can then be identified.

Confidence Intervals and Hypothesis Testing using Euclidean Distance Matrix Analysis (EDMA)

Lele and Richtsmeier (1995) introduced procedures to obtain confidence intervals for elements of the form difference matrix. They proposed two approaches. One is model based using a Gaussian model for the random error and employs Monte Carlo techniques to estimate parameters from the samples directly. The second is a bootstrap method, which does not depend on the assumption of an underlying Gaussian model. The former is appropriate if the underlying distribution is believed to be Gaussian. If this is not the case then the bootstrap method is more appropriate provided samples sizes are sufficient. They also propose two methods for hypothesis testing. Essentially, the null hypothesis states that the shapes are similar. Hypothesis procedures can be employed to test if differences in mean form or shape infer that the samples were drawn from different shape or form populations using the SDM and FDM (form difference matrix). EDMA-I (Lele, 1991) compares two samples of forms, treating one as a baseline group, and so is considered a one sided test. There is an assumption that the variances of the two populations are equal. The procedure calculates the observed ratio T_{obs} of the largest and smallest values of the form difference matrix for the mean shapes of the

two samples, under the assumption that one group's form is purely a scaled version of the other.

$$T_{obs} = \frac{\hat{d}_{max}}{\hat{d}_{min}},$$

where d is an element of the FDM. T_{obs} can take values greater than or equal to unity (the baseline group is chosen to ensure this). If the samples are from the same shape population (any scale differences between the two groups are removed when taking the ratio) then T_{obs} should be close to unity. An estimate for the null hypothesis distribution of T_{obs}, stating that the samples are from the same population is obtained using bootstrap sampling. The null hypothesis is rejected if T_{obs} falls in the upper $\alpha\%$ tail.

EDMA-II (Lele and Cole, 1996) is a two-way test that does not require one group to be chosen as a baseline. Also, the test does not require the population variance of the two groups to be equal. The SDM is used to construct the test statistic, which is the absolute value of the element with the largest absolute value. The mean shape matrices are dependent on the choice of scaling factor. A Gaussian model is assumed and Monte Carlo parametric bootstrap samples obtained for the test statistic. The Z statistic is calculated for each sample. The sample is ordered and used to accept or reject the null hypothesis if zero is within the chosen $100(1-\alpha)$ % confidence interval. Lele and Cole point out that if the maximum and minimum absolute values of the $SDM(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are similar in magnitude the bootstrap histogram may be bimodal, in which case the null hypothesis should be rejected.

It is emphasized that a simple test between mean shapes is not what is usually required. Rather, most interest is usually in those differences within the landmark configurations. Foe instance, where do landmarks differ significantly or where are they similar. This is not captured by simply testing for an overall difference in shape.

Recovering coordinates from a Euclidean Distance Matrix (EDM)

There is a unique form representation for every landmark configuration,

$$\mathbf{X} \mapsto FM(\mathbf{X})$$

which is a single point in the $\frac{k(k-1)}{2}$ dimensional form space. However, there is not a unique inverse mapping back to the landmark configuration, which is termed an orbit. A single representative configuration, termed an icon, can be found using metric scaling, this ensuring the squared Euclidean distance matrix is positive semi-definite. For any $\mathbf{X}$, k points in $\mathbb{R}^m$, the matrix of Euclidean distances $\mathbf{D} = [d_{rs}^2]$, where $d_{rs}^2 = \|x_r - x_s\|$ can be made positive semi-definite, $\bar{\mathbf{D}}$

$$\bar{\mathbf{D}} = -\frac{1}{2}\left[\mathbf{D} - \frac{(\mathbf{D}\underline{\mathbf{1}})\underline{\mathbf{1}}'}{k} - \frac{\underline{\mathbf{1}}(D\underline{\mathbf{1}})'}{k} + \frac{\underline{\mathbf{1}}'\mathbf{D}\underline{\mathbf{1}}}{k^2}\right].$$

Also, given a symmetric $\mathbf{D}$ with positive semi-definite $\bar{\mathbf{D}}$, a configuration of points can be found in $\mathbb{R}^{(k-1)}$ such that $\mathbf{D} = [d_{rs}^2]$. A necessary and sufficient condition for a $k \times k$ matrix $\mathbf{D}$ to be a squared distance matrix is that

$\mathbf{w}'\mathbf{D}\,\mathbf{w} \leq 0 \quad \forall\, \mathbf{w}'\underline{\mathbf{1}} = 0,$

N.B. $\text{rank}(\bar{\mathbf{D}}) = \text{rank}(\mathbf{X} - \underline{\mathbf{1}}(\mathbf{X}\underline{\mathbf{1}}))$.

The positive semi-definite matrix $\bar{\mathbf{D}}$ will possess m non zero eigenvalues and corresponding eigenvectors. These can now be used to form a representative of the orbit of $FM(\mathbf{X})$

$$\mathbf{A}_d = [\sqrt{\lambda_1}\mathbf{v}_1, \ldots, \sqrt{\lambda_1}\mathbf{v}_m].$$

then, $FM(\mathbf{A}_d) = FM(\mathbf{X})$.

1.4.4 Problems with Procrustes Superimposition

If Procrustes superimposition is applied to a sample of shapes, under the assumption that they have a Gaussian distribution, then this does not eliminate the Euclidean nuisance parameters, rather they are constrained. This has the effect of making the Procrustes mean and variance-covariance matrix inconsistent. Lele (1993) has shown that in the limit they will not converge to the true population values. Kent and Mardia (1997) investigated this and found that the Procrustes estimator of shape is consistent only under the assumption of an isotropic error. Also, PCA based on Procrustes residuals can be misleading when variance is estimated inconsistently (Procrustes residuals are approximate tangent coordinates). The landmarks farthest away from the centroid are matched closest at the cost of those closer to the centroid. This means that landmarks close to the centroid have inflated variance, while those further away have deflated variance estimates. The effect is most pronounced when the landmark configuration is not symmetrical around the centroid. So, for Procrustes superimposition, variability is driven by the method and not according to the natural variability of the configurations.

In Chapter 4, an EDM approach is taken to eliminate the problems discussed with using shape coordinate systems.

Chapter 2

Simple Component Analysis

2.1 Introduction

In the 1930's Thurstone pioneered the definition of simple structure in factor analysis (for example, Thurstone, 1931). In order to interpret factors, it is desirable to give the simplest explanation in terms of the observed data. Often no prior labels are attached to factors and the analysis is used in an exploratory fashion to extract some lower dimensional definitions from the observed variables. For instance, in a consumer panel test a derived factor may subsequently be labeled as representing *health* after inspection of its weights. Thurstone proposed the following criteria.

1. Factors should have simple structure to explain the correlation between observed variables, with as many near zero and high weights being present.
2. An observed variable should be weighted heavily on one (or a small number) of factors.
3. Each factor should have only a few variables highly weighted on it, and therefore factors are specifically related to clusters of interdependent observed variables.
4. Ideally factors should isolate those variables that respond to the same causal influence.

PCA (page 3) is a simplifying technique which seeks the true dimensionality of a data set. A PCA is an optimal linear procedure in that it finds axes in the direction of the maximum variation in the data and produces scores which are uncorrelated with each other. Any other linear procedure is sub-optimal to a PCA in this sense. However, a consequence is that interpretation of its components is compromised and Thurstone's criteria are not met.

2.2 The Interpretation of PCA

Table 2.1 shows the loading vectors from a PCA on data collected from another deodorant test panel. A subset of twelve variables were chosen for illustration. This

	L1	L2	L3	L4	L5
Fresh deodorant	-0.287	0.366	-0.212	0.225	-0.008
Fresh on application	0.334	-0.313	0.175	-0.227	-0.201
Pleasant on skin during application	0.339	-0.058	0.122	0.018	-0.433
Gentle on skin	0.324	-0.038	0.008	0.112	-0.514
Did not mark clothes	0.236	0.062	-0.642	-0.136	0.005
Did not leave white marks	0.240	0.048	-0.635	-0.163	-0.002
Feeling fresh all day	0.324	-0.240	-0.018	0.404	0.424
Feeling confident	0.340	-0.236	-0.011	0.398	0.331
Sticky	-0.292	-0.356	-0.159	0.341	-0.146
Greasy	-0.271	-0.365	-0.181	0.273	-0.214
Wetness	-0.271	-0.367	-0.190	0.016	-0.254
Coldness	-0.141	-0.501	-0.021	-0.572	0.295
Variance explained	5.45	1.46	1.36	0.89	0.77

Table 2.1: Principal component analysis of deodorant data, showing the first five loading vectors. Many small weights are present which together account for a large proportion of the variance. However, these weights make interpretation difficult.

example illustrates that there can be many loadings that although small may together explain a significant amount of variation in the data. However, it is difficult to interpret the importance of each. For example, on L3 there are two values of magnitude 0.6, five values of approximate magnitude 0.2 and then five less than 0.2. These principal component loadings do not conform to the criteria of Thurstone. Other researchers have proposed methods to improve the interpretation of PCA. These are outlined in the following sections and a new method will be discussed in this chapter. The goal is to replace principal components by a system which is more interpretable. In order to do this some of the optimal features of a PCA must be sacrificed; less variability will be extracted with each component and they will be correlated with one another, or axes may become oblique. If the loss is small then it is attractive to use a more interpretable system. To address this trade off between optimality and simplicity, a number of approaches have been adopted.

2.2.1 Rotation to Simple Structure

A simpler solution can be sought starting from an optimal system of components obtained from a PCA. A rotation of a subset of the principal components to a simpler structure will conserve the total variance explained by them, but it becomes more spread between the components with either a loss of orthonormality or the introduc-

tion of some correlation between the scores. The varimax procedure, is one of the earliest methods developed by Kaiser (1958) and forces solutions with a small number of large weights on each component (page 22). Unfortunately, many small weights can still be present and although it is tempting to interpret the rotated components in terms of their large weights, the non-informative weights can still be important. Table 2.2 shows the rotated loadings vectors for the deodorant example.

	RL1	RL2	RL3	RL4	RL5
Fresh deodorant	-0.253	0.025	-0.076	0.101	-0.327
Fresh on application	0.412	0.026	-0.103	-0.052	0.203
Pleasant on skin during appl	0.443	0.099	-0.186	-0.035	-0.195
Gentle on skin	0.433	0.200	-0.170	0.036	-0.328
Did not mark clothes	-0.066	0.038	-0.074	0.568	0.046
Did not leave white marks	-0.074	0.026	-0.067	0.566	0.063
Feeling fresh all day	-0.197	0.054	0.642	-0.051	-0.008
Feeling confident	-0.132	0.079	0.587	-0.052	-0.052
Sticky	0.055	0.487	0.146	-0.012	-0.152
Greasy	0.125	0.484	0.061	0.024	-0.099
Wetness	0.158	0.404	-0.070	0.081	0.056
Coldness	0.009	-0.056	-0.018	0.067	0.805
Variance explained	2.71	2.27	1.91	1.87	1.15

Table 2.2: The first five rotated loading vectors for the deodorant example. There are many more near zero loadings. Although there is now obvious differentiation, for instance RL2 where eight of the loadings are near zero, together they still account for a significant proportion of the variation.

Practitioners often take loading vectors from a principal component analysis and select subsets of weights by setting those which appear small to zero. Unfortunately, this can lead to poor approximations and incorrect interpretation as illustrated by Cadima and Jolliffe (1995). They show that interpreting loading vectors based on either the absolute magnitude of loadings or their correlation with a given component can give incorrect interpretation of their importance.

2.2.2 The Simplified Component Technique

An alternative to rotated principal components was proposed by Jolliffe and Uddin (2000) who formulated the problem into a single step by the addition of a penalty constraint to the optimization problem that favours simplicity. This simplified component technique (SCoT) is formulated by combining the sequential extraction of principal components with a simplicity constraint such as varimax up front. In so doing, it implicitly manages the trade off between the maximal variance properties of PCA and simplicity. The method seeks $\{\mathbf{a}_k\}$ such that

$$\text{var}\left(\mathbf{a}_k'\mathbf{x}\right) + \phi F\left(\mathbf{a}_k\right)$$

is maximized, subject to $\mathbf{a}_k'\mathbf{a}_k = 1$ and $\mathbf{a}_j'\mathbf{a}_k = 0$ or $\mathbf{a}_j'\mathbf{\Sigma}\mathbf{a}_k = 0 \; \forall \; j, k \, (j \neq k)$, and where F is a simplicity function such as varimax and ϕ is a simplicity/complexity parameter. Successive components are found to be either orthogonal or uncorrelated.

SCoT is not equivalent to a PCA followed by the rotation of a subset of components as the axes remain within the subspace of the chosen components whereas with SCoT they will not. SCoT finds components sequentially and so those previously found will not change. However, a rotation of additional PCs may change the nature of all. SCoT is a more complex optimization problem and finding orthogonal and uncorrelated components are not equivalent unlike with PCA. As the optimization criteria is more complex the tendency is to find local optima. Jolliffe et al. (2003) replace the explicit simplifying criterion of SCoT with the LASSO (least absolute shrinkage and selection operator), Tibshirani (1996), which usually produces some exact zero loadings. In multiple regression, the LASSO imposes an additional restriction on the coefficients.

$$\sum_{j=1}^{p} |\beta_j| \leq t,$$

for some tuning parameter t. For suitable values of t, this constraint has the property that some of the coefficients in the regression will be exactly zero. The simplified component technique using the LASSO (SCotLASS), has the simplifying constraint

$$\sum_{j=1}^{p} |a_{kj}| \leq t,$$

where a_{kj} is the jth element of the kth loading vector. Like other implementations of SCoT, SCoTLASS is a non-trivial optimization problem and suffers from the same problems. Zou et al. (2006) point out that SCoTLASS lacks guidance for the choice of t and that the high computational cost is due to the optimization problem being non convex. Also the solutions are not sparse enough when requiring a high percentage of explained variance. Their method Sparse PCA (SPCA) treats PCA as a ridge regression problem employing the LASSO. As a procedure, SPCA enjoys advantages in several aspects, including computational efficiency, high explained variance and ability of identifying important variables. A unified and efficient algorithm has been proposed to realize SPCA for both regular multivariate data and gene expression arrays where the number of variables exceeds the number of observations. It allows flexible control of the sparse structure of the resulting loadings. The SPCA criterion gives exact PCA results when its sparsity (lasso) penalty term vanishes.

2.2.3 Simple Systems of Components

The SCoT and SPCA do not restrict the nonzero loadings to a discrete set of values. If there is a definition of simplicity then optimal solutions can be sought from simple

systems of components. These will be truly simple as there is no implicit or explicit requirement to decide on the importance of particular loadings on the components. Vines (2000) used simplicity performing transformations to search for loading vectors which are proportional to integers. These are obtained using constrained rotations and rescaling with pairs of variables. This approach requires many searches over all pairs of variables ($p(p-1)/2$ searches). If in addition to the conditions defined by Thurstone, simplicity is extended to the form of the weights so that it is not required to decide if a weight is important then interpretation of individual loadings is trivial; either anti correlated, correlated or zero. This can be achieved by restricting the weights to be scaled integers belonging to the set $\{-1, 0, 1\}$. This particular set of integers was termed Hausman weights by Choulakian et al. (2006c) after Hausman (1982) published a branch and bound algorithm that finds the solution which maximizes the variance but subject to the integer constraint on the weights. In its original form strict orthogonality was not imposed, however this was forced by inclusion of a suitable constraint (DeSarbo and Hausman, 2005). This was shown to be better than an exhaustive search. One of the problems with this approach is the computational overhead associated with the branch and bound algorithm and consequently it does not scale well. D'Aspremont et al. (2004), approximate the covariance matrix by a rank-one matrix using semi-definite programming to decompose it into sparse factors, minimizing.

$$\left\|\boldsymbol{\Sigma} - \mathbf{a}\mathbf{a}'\right\|$$

subject to $\text{Card}(\mathbf{a}) \leq k$. Sparseness is introduced by solving a relaxed optimization problem where the number of non zero loadings is introduced as a cardinality constraint k. Recently Witten et al. (2009) consider a penalized matrix decomposition to find sparse principal components while Johnstone and Lu (2009) consider their consistency.

A simple system of components using Hausman weights was obtained by the enumeration of all possible configurations for the small deodorant example. The enumerated solution is shown in Table 2.3 and can be compared with the principal component solution in Table 2.1. The enumerated solutions are found that maximize the objective function (2.13), and where more than one solution is found the one with the smallest loss is preferred.

Table 2.3 shows the first five unscaled simple components, obtained for the example deodorant data introduced earlier in this section. These were calculated by the sequential enumeration of each component. The weights are Hausman weights. In this case all possible patterns of {-1, 0, 1} are considered for each component and subsequent components are found subject to those previously found. The coefficients 0, ± 1, are adjusted to give a component length of unity. The objective function used penalizes non-orthogonal component loading vectors. This enumeration is not guaranteed to give the global optimum, which can only be guaranteed if all components are enumerated

	E1	E2	E3	E4	E5
Fresh deodorant	1	1	0	1	0
Fresh on application	-1	-1	0	-1	-1
Pleasant on skin during application	-1	-1	0	0	-1
Gentle on skin	-1	0	0	1	-1
Did not mark clothes	-1	1	1	0	1
Did not leave white marks	-1	1	1	-1	0
Feeling fresh all day	-1	-1	0	1	1
Feeling confident	-1	-1	0	1	1
Sticky	1	-1	1	1	0
Greasy	1	-1	1	0	-1
Wetness	1	-1	0	-1	1
Coldness	0	-1	0	-1	1
Variance explained	5.29	1.37	1.17	0.81	0.77

Table 2.3: The unscaled enumerated simple components for the deodorant example

simultaneously. Similarities can be seen between the principal components and the simple components. The smallest loading on L1 becomes zero on E1. The remaining coefficients on E1 are ± 1, but opposite signs to L1, but these can be reversed. Corresponding principal components and simple components explain roughly the same variation. The loadings are clear to interpret, for example, previously L3 had five loadings that were less than 0.2. The enumerated component E3 now is sparse and its interpretation is a weighted average representing the propensity of the deodorant not to mark clothes and to feel greasy and sticky. The optimization criterion used for the enumeration is an analogue of PCA, the loadings obtained after a rotation such as the varimax may differentiate better due to the nature of the varimax criterion. However, any optimization criterion can be used with the enumeration. Another advantage of this approach over PCA is that the loadings are always Hausman weights. For larger problems the PCA loadings will become smaller and harder to differentiate, due to the length constraint of unity placed on the principal component.

Techniques that approximate principal components

Recently Chipman and Gu (2005) extended the earlier work of Vines to produce loading vectors proportional to vectors of small integers, but individually close to their PC counterparts (measured by their angle) and pairwise orthogonal. Simplicity is considered in the broad sense of the appearance of useful structure in the loadings. Rather than solve an explicitly constrained optimization, Chipman and Gu introduce sparseness and homogeneity into the loadings. A homogeneity constraint causes each weight to be proportional to ± 1 or 0, which are Hausman weights. If the weights are constrained to sum to zero this will favour *contrasts*. There are $3^p/2$ possible values for

the loading vector $\mathbf{a}_i$. To find the best $\mathbf{a}_i$ the angle to the ith principal component is minimized subject to the weights taking values of zero or $\pm c$, where c is determined by the length of the vector (2.1). This is tested by minimizing $\arccos(\mathbf{u}_i'\mathbf{a}_i)$ or equivalently maximizing $\mathbf{u}_i'\mathbf{a}_i$; $\mathbf{u}_i$ is the ith principal component direction. The original principal component axis $\mathbf{u}_i$, is truncated by taking the largest absolute weights for increasing length k and replacing with 1 or -1. For illustration, in the example taken from their article the component direction $\mathbf{u}_i = (0.41, -0.03, -0.42, 0.81)$ is approximated with k = 1 to 4 non-zero elements,

$$\begin{aligned}
\mathbf{a}_i &= (\ 0,\ 0,\ 0,\ 1) && k=1 \\
\mathbf{a}_i &= \frac{1}{\sqrt{2}}(\ 0,\ 0,-1,\ 1) && k=2 \\
\mathbf{a}_i &= \frac{1}{\sqrt{3}}(\ 1,\ 0,-1,\ 1) && k=3 \\
\mathbf{a}_i &= \frac{1}{2}(\ 1,-1,-1,\ 1) && k=4,
\end{aligned} \tag{2.1}$$

in which case $[1\ 0\ -1\ 1]/\sqrt{3}$ is closest to $\mathbf{u}_i$ with an angle of 18.8 degrees. Sparse solutions are desirable as this separates out interdependent variables onto separate components. Sparsity constraints are implemented using one of the following penalties:

$$C1 = \frac{\theta}{(\pi/2)} + \frac{\eta k}{p}$$

or

$$C2 = (p-k)\,(\cos\theta)^{\eta},$$

where θ is the angle between the $\mathbf{a}_i$ and $\mathbf{u}_i$ and η is a tuning parameter. In the case of $C1$, as η increases the components become more sparse. For $C2$, the solutions will become less sparse. A stepwise algorithm is proposed which involves finding each interpretable component sequentially, by deflating the covariance matrix after each component is found. Then the next component is approximated to the largest eigenvector corresponding to the largest eigenvalue of the residual covariance matrix.

The algorithm scales linearly, but cannot guarantee explaining maximal variance. According to Chipman and Gu it performs similarly to Vines approach except when Vines loading vectors are proportional to large integers. As this approach approximates principal components by ignoring small weights, it is susceptible to the problems discussed earlier in Section 1.1.3, as many combinations of variables are ignored and there is an explicit assumption that the largest weights are the most important. To ensure that the components do not diverge too far from the principal components, deflation of the covariance matrix is included. Then, subsequent components match the largest eigenvector of the current deflated covariance matrix. Anaya-Izquierdo et al. (2008) discuss a method to find simple components that approximate the principal components using an approach similar to Chipman and Gu (2005). However, this differs in

that the algorithm ensures that the components are orthogonal by finding a subspace that is guaranteed to be orthogonal. However, the weights cannot always be Hausman, as there is no guarantee that solutions with purely Hausman weights exist in a strictly orthogonal subspace.

Hausman Principal Components

Choulakian (2001, 2003, 2005, 2006a), Choulakian et al. (2006b) describe a two step method to obtain Hausman principal components (HPC), using the centroid method originally proposed by Burt (1917), and further developed by Thurstone (1931). The centroid method gives approximate principal components with scaled integer weights belonging to $\{1, -1\}$. Weights are then chosen to be zeroed based on finding the subset that best approximates the original principal component, consistent with the ideas presented in Cadima and Jolliffe (1995). The first stage solves the problem

$$\max \ \mathbf{a}\mathbf{Z}\mathbf{Z}'\mathbf{a} \text{ subject to } a_i \in \{-1, 1\} \text{ for } i = 1, \ldots, n$$

where $\mathbf{Z}$ is a $n \times p$ standardized data matrix and $\mathbf{a} \in \mathbb{R}^n$. This takes $(2^p - 1)$ centroid PCAs which is NP hard and produces loadings of 1 or -1. The second stage is combinatorial taking pairs of variables and assessing the change in the average variance after their deletion from the centroid solution of stage one. A global solution is not guaranteed.

Clustering Approach

Rousson and Gasser (2004) proposed a method in two steps to find interpretable components. These are suboptimal sacrificing some of the information and inducing correlation. The structure in the correlation matrix is exploited by clustering the variables into multiple blocks that are not too correlated. The correlation between components should remain low if they are to be interpreted independently. The method allows more than one block component, unlike PCA which usually defines only one. The block components are found using agglomerative clustering. Once the blocks are found contrasts are derived iteratively. The contrasts sum to zero and are found within blocks by regressing the already extracted components onto the original variables and then taking the first principal component of the residual correlation matrix. However, these are not necessarily orthogonal. As each contrast is found it is simplified by counting

the number of positive and negative PCA loadings. For example,

PC Loading	Simplified Loading
0.027	2
−0.388	−4
0.032	2
0.579	2
0.164	2
−0.697	−4

so that the principal component loadings in column one would become the contrast in column two. In this case there are four positive loadings and so the two negative loadings are replaced by the integer -4 and the positive loadings by the integer 2.

2.2.4 Problem Complexity

In an article by Ferrez et al. (2005) the complexity of solving a fixed rank quadratic optimization with discrete values restricted to be in the set $\{0, 1\}$ shows the difficulty involved in solving this simpler problem. Due to its combinatorial nature it is not open to dynamic programming (the centroid method provides an heuristic method to find approximate principal components). Other direct optimization approaches such as branch and bound and semi-definite programming (SDP) do not scale well. As discussed other approaches either use a penalized form of optimization but still find real valued loadings or approximate principal components with discrete values. In some of the latter cases consideration is given to the problems explored by Cadima and Jolliffe (1995, 2001), for example Hausman principal components Choulakian et al. (2006c). Simple component analysis Rousson and Gasser (2004) retain the sign of the principal component loadings within contrasts.

Heuristic strategies such as a *greedy search* have been found to find good solutions in tractable time for some problems. A greedy algorithm is any algorithm that follows the problem solving meta-heuristic of making the locally optimal choice at each stage with the hope of finding the global optimum (see, Cormen et al., 2001, for an introductory text). One of the goals of this thesis is to find tractable simple solutions for large data sets, containing potentially hundreds of variables. A greedy approach may allow good solutions to be found in the sense of maximizing variance and can be combined with other structure simplifying approaches.

2.3 Finding Simple Components

The problem of finding components that sequentially maximize the variance explained can be posed in a different manner to that described in Chapter 1. Components are

constructed from data as weighted sums of the variables. If $\mathbf{X}$ is a $n \times p$ centred data matrix (n observations on p variables) then the projection of $\mathbf{X}$ onto a vector $\mathbf{w}_1$ will form scores $\mathbf{y}_1$ on the axis defined by $\mathbf{w}_1$,

$$\mathbf{y}_1 = \mathbf{X}\frac{\mathbf{w}_1}{\|\mathbf{w}_1\|}. \tag{2.2}$$

Principal components finds the axis $\mathbf{w}_1$ such that data projected onto it accounts for the maximum amount of variation, so that

$$\text{var}\,(\mathbf{y}_1) = \frac{\mathbf{w}_1'\mathbf{X}'\mathbf{X}\mathbf{w}_1}{\mathbf{w}_1'\mathbf{w}_1} \tag{2.3}$$

is maximized.

$\mathbf{X}'\mathbf{X}$ is related to the sample covariance matrix $\mathbf{S}$, and is symmetric and positive semi definite ($\mathbf{x}'\mathbf{S}\mathbf{x} \geq 0 \;\forall\; \mathbf{x}$). The $\mathbf{w}_1$ that maximizes the quotient is the eigenvector associated with the largest eigenvalue. Then, under the condition of orthogonality between axes, the next axis $\mathbf{w}_2$ explains the most variation after that, subject to $\mathbf{w}_2'\mathbf{w}_1 = 0$, and is the eigenvector associated with the second largest eigenvalue and so on for $\mathbf{w}_3, \ldots, \mathbf{w}_p$.

Mathematically, the problem of finding Hausman weighted vectors to explain the maximum amount of variation subject to some penalty is

$$\max_w \; \frac{\mathbf{w}'\mathbf{S}\mathbf{w}}{\mathbf{w}'\mathbf{w}} - \tau \times \text{loss} \tag{2.4}$$

or

$$\max_w \; \frac{\mathbf{w}'\mathbf{S}^2\mathbf{w}}{\mathbf{w}'\mathbf{S}\mathbf{w}} - \tau \times \text{loss}$$

subject to $w_i \in \{-1, 0, 1\}$, and τ is a penalty parameter whose value also has to be chosen. The loss can favour components which are orthogonal, for example

$$\text{loss} = \sum_{i=1}^{k} \left(\frac{\mathbf{a}_i'\mathbf{w}}{\|\mathbf{a}_i\| \, \|\mathbf{w}\|} \right)^2 \tag{2.5}$$

or have uncorrelated scores

$$\text{loss} = \sum_{i=1}^{k} \text{corr} \left(\frac{\mathbf{X}\mathbf{a}_i}{\|\mathbf{a}_i\|}, \frac{\mathbf{X}\mathbf{w}}{\|\mathbf{w}\|} \right)^2 \tag{2.6}$$

where $\mathbf{a}_i$ is a previously found weight vector and the square is taken to ensure the loss is greater than zero. Another interesting case is when solutions are sought where the components have maximum correlation with each, other or within a subset. The

utility of orthogonal correlated components is investigated in Chapter 3 in the context of rotating principal components. In the context of simple components a modified loss could be used. Unfortunately, this is not accessible to techniques such as dynamic programming. The complexity of solving the simpler unconstrained convex optimization problem was mentioned earlier

$$\max_{w} \ \mathbf{w}'\mathbf{S}\mathbf{w} \text{ subject to } w_i \in \{0, 1\}. \tag{2.7}$$

In essence, to find a globally maximum solution previously found, weights may require further updating. Full enumeration is impracticable for large data sets; for example a data set with twenty variables would require $\frac{3^{20}}{2}$ comparisons to find the first component. Greedy algorithms can provide computationally tractable solutions, which although not globally optimum, may still provide very good solutions.

2.4 A New Greedy Algorithm to Find Simple Components

An algorithm is termed greedy if at an iteration it only considers the best local solution and does not reconsider any previous decision. Full enumeration of a single component consisting of p loadings requires the comparison of $(3^p - 1)/2$ cases. This is feasible for small problems but not tractable for larger numbers of variables. Dynamic programming approaches are not appropriate as previously considered cases can become important after considering subsequent cases.

It is proposed that variable tuples are considered together and these are fully enumerated. If all combinations of k tuples are taken then solutions are found which although not guaranteed to be globally optimum may provide good solutions compared to the optimal properties of PCA. There are $p!/(p-k)!k!$ combinations to consider on a component and each k tuple requires 3^k patterns to be considered. As a simple illustration taking pairwise tuples, the patterns to consider for each pair (i, j) of variable loadings are shown in Table 2.4

The idea is presented in pseudo-code as **Algorithm 1**. To initialize a random or homogeneous vector is used. Alternatively the sign function can be applied to threshold the principal component, where any weight with absolute value less than a chosen threshold is zeroed. Combinations are considered at random from the $p!/(p-k)!k!$ possibilities. To extract all p components there are

$$p3^k \frac{p!}{(p-k)!k!}$$

combinations to evaluate. To be useful the number of comparisons should be consider-

ith	jth
0	0
0	1
0	-1
1	0
1	1
1	-1
-1	0
-1	1
-1	-1

Table 2.4: Each row is the pair of simple unscaled loadings to consider for the selected pair of variables

ably less than full enumeration, so that

$$3^k \frac{p!}{(p-k)!k!} << \frac{3^p - 1}{2}$$

If k is kept small, less than 3 or 4 then solutions for large data sets become feasible. In fact k must be less than $0.4p$ to be a better option than enumeration. For small k the algorithm may be repeated starting from the best solution found to date. Once a simple component has been found the next is found subject to maximizing the objective described earlier in equation (2.4).

If strict orthogonality is required, then a subsequent simple component could be sought with the extra constraint that it is in the $p - q$ subspace orthogonal to the q components already found. However, a simple component consisting of Hausman weights is not guaranteed to be in an orthogonal subspace. Alternatively, in a similar way to *principal component regression* the simple components can be regressed onto an independent variable, in which case it is desirable for them to be uncorrelated.

Finding principal components sequentially is optimal as the global solution is the solution to an eigenvalue problem, however for simple components this is not the case. Consequently, better solutions may be obtained by considering groups of simple components simultaneously. However, the search space is expanded. If q is the number of components to consider together and k the tuple size to consider across the q components, then the number of comparisons is

$$3^{kq} \frac{p!}{(p-k)!k!} \tag{2.8}$$

as each k tuple across q components requires 3^{kq} comparisons. Sequentially there are

$$q3^k \frac{p!}{(p-k)!k!} \tag{2.9}$$

comparisons. So finding components simultaneously requires

$$\frac{p!\left(3^{qk} - q3^k\right)}{(p-k)!k!}$$

extra comparisons or

$$\frac{3^{k(q-1)}}{q}$$

times as many. For example to find 3 simple components simultaneously for a problem involving 20 variables, simultaneous enumeration requires $3^{20\times 3} \approx 4\times 10^{28}$ comparisons; simultaneous simple components taking pairs requires 138,510 and sequential simple components requires 5,130. Sequential enumeration would require $3 \times 3^{20}/2 \approx 5^9$ comparisons. Clearly full simultaneous enumeration is likely to be impractical. Sequential enumeration would be practical for small problems and simultaneous simple components would be tractable for moderate size problems. However a sequential simple component approach may find solutions to large problems so long as these solutions are close to principal components in terms of the variance they explain and their orthogonality or correlation.

Input: Covariance matrix S $(p \times p)$. The number of components to find. The number of elements to consider simultaneously k. The penalty weight λ
Output: Simple Component set C $(q \times p)$
repeat
  repeat
    Initialize the simple component { Random, First Principal Component, Vector of Ones}
    Apply threshold to make it simple
    Reorganize the vector and covariance into fixed and variable parts to reduce computation
    forall the *possible patterns (see Table 2.4) for k variables* **do**
      Evaluate the loss in Equation (2.13) and keep the best solution
    end
  until *All k combinations of p variables have been evaluated*;
until *All q components are extracted*;

Algorithm 1: Outline of the basic greedy search approach to find simple components.

Returning to the example in Section 2.2, Table 2.5 shows the simple components found for $k = 3$ using a squared orthogonal penalty. This can be compared with those found by enumeration, Table 2.3. Notice that the first three components describe identical axes. However, the enumerated solutions explain slightly more variation overall. This is still encouraging as the number of comparisons is substantially less than full enumeration. Full enumeration of q components requires $q\frac{3^p}{2}$ comparisons. For these data this equates to $1,328,600$ comparisons by enumeration compared to $29,700$ using the simple component search. In this case the simple component solution is more sparse than after sequential enumeration. For instance SC4 can be easily interpreted as a contrast between the freshness on application of the deodorant, compared to the freshness after application and feeling confidence. SC5 contrasts freshness with gentle and pleasant. The performance of the sequential simple component algorithm is ex-

	SC1	SC2	SC3	SC4	SC5
Fresh deodorant	1	1	0	1	1
Fresh on application	-1	-1	0	-1	0
Pleasant on skin during application	-1	-1	0	0	1
Gentle on skin	-1	0	0	0	1
Did not mark clothes	-1	1	1	0	0
Did not leave white marks	-1	1	1	0	0
Feeling fresh all day	-1	-1	0	1	-1
Feeling confident	-1	-1	0	1	0
Sticky	1	-1	1	0	0
Greasy	1	-1	1	0	0
Wetness	1	-1	0	0	0
Coldness	0	-1	0	0	1
Variance explained	5.29	1.38	1.17	0.77	0.6

Table 2.5: The unscaled components found with the simple component algorithm ($k = 3$, squared orthogonal penalty).

plored in the next section. If good solutions are found in polynomial time then there may be value for large problems.

2.5 Assessing the Quality of Solutions

This section considers the quality of solutions found when finding simple components. The following are considered when comparing component sets,

1. How much of the variance is explained compared to PCA, which is sequentially optimal
2. How correlated are the resultant simple component set
3. Is there a loss of orthogonality
4. The reconstruction error when re-estimating the data matrix from a simple components set compared to PCA.
5. Interpretable components cannot be guaranteed, however simplicity as measured by sparsity and contrasts will often lead to more interpretable components.

How to assess the quality of a component set, which can be correlated and non-orthogonal, is reviewed by Gervini and Rousson (2004). They maintain that any measure must possess *generality* and *uniqueness*. Generality refers to a criteria being applicable to a broad range of unit norm components with linearly independent loading vectors. Consequently, any criterion which depends on the loading vectors being orthogonal or uncorrelated will violate generality. To show uniqueness the variance

of the measure must be maximized by the first q principal components, so that any correlation or loss of orthogonality between a set of loading vectors is penalized. The best linear predictor (BLP), correlation of projections (CP), generalized variance (GV) and RV-coefficient (RV) do not differentiate adequately between component sets based on these criteria. For example, the BLP only depends on the subspace spanned by the loading vectors and so is invariant after a rotation or any full rank transformation. The GV is maximized by the principal components, without assuming the loading vectors are orthogonal or uncorrelated and so satisfies generality. However, it is invariant under rotation and so is not unique.

2.5.1 Metrics

The following measures were chosen to quantify the performance of the simple component algorithm.

1. Gervini and Rousson develop criteria to evaluate the loss of orthogonality and increase in correlation compared to a subset of principal components. The corrected sum of the variance (CSV) is defined as

$$\begin{aligned} \mathrm{CSV}(\mathbf{A}_\mathrm{q}) &= \frac{\mathrm{trace}\left(\mathbf{A}_q'\mathbf{S}\mathbf{A}_q\right)}{\mathrm{trace}\left(\mathbf{\Delta}_q\right)} \\ &- \frac{\sum_{k=2}^{q}\left(\mathbf{a}_k'\mathbf{S}\mathbf{A}_{(k-1)}\left(\mathbf{A}_{(k-1)}'\mathbf{S}\mathbf{A}_{(k-1)}\right)^{-1}\mathbf{A}_{(k-1)}'\mathbf{S}\mathbf{a}_k\right)}{\mathrm{trace}\left(\mathbf{\Delta}_q\right)} \\ &= Opt_1 - Opt_2, \end{aligned} \tag{2.10}$$

   where $\mathbf{A}_q$ is the matrix whose columns are the weight vectors of the q chosen components and $\mathbf{\Delta}_q$ is a diagonal matrix of eigenvalues from the corresponding q principal components.

   It captures the contribution of adding successive components to the system by summing the residual variance. If the system is correlated then the second term (Opt_2) will be non-zero. In fact, the second term will be zero if and only if the components are the principal components, and then the first term (Opt_1) will be unity. Thus, $\mathrm{CSV}(\mathbf{A}_\mathrm{q}) \leq 1$ in general, with $\mathrm{CSV}(\mathbf{A}_\mathrm{q}) = 1 \iff \mathbf{A}_q = \mathbf{U}_q$, where $\mathbf{U}_q$ is a set of q principal component loading vectors.

   The CSV and its components Opt_1 and Opt_2 can be used to compare systems for their correlation and orthogonality. The value of $Opt_2 = Opt_1 - CSV\left(\mathbf{A}_q\right)$ is a measure of the correlation between components. One draw back is that it is not symmetric in the q components, so the value of $\mathrm{CSV}(\mathbf{A}_\mathrm{q})$ depends on the

ordering of the components. For consistency components are ordered in terms of the descending variance explained.

2. A simple measure of orthogonality of a set of components $\mathbf{A}$ is $\sum_{j>i}(\mathbf{a}_i'\mathbf{a}_j)^2$. An orthogonal set has a value of zero. For equal numbers of components the value can be used to rank the solutions.

3. Reconstruction error
   The principal components will give minimal reconstruction error. If a subset of component loading vectors are taken $\mathbf{Y}_k = \mathbf{X}\mathbf{A}_k$, then the reconstructed data $\hat{\mathbf{X}} = \mathbf{Y}_k(\mathbf{Y}_k'\mathbf{Y}_k)^{-1}\mathbf{Y}_k'\mathbf{X} = \mathbf{Y}_k\hat{\mathbf{P}}$. The error is then $\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{Y}_k\hat{\mathbf{P}}$. The trace of the variance of $\mathbf{E}$ gives the unexplained variance when reconstructing the original data using the chosen k axes of $\mathbf{A}$.

4. Recovery of a known structure would be another possibility.

The solutions are assessed using these metrics.The principal components are optimal in certain ways. However, another benchmark is the full enumeration of simple components by considering every possible combination using the same optimization criterion as for the simple components. In addition to PCA and full enumeration comparisons are made against a random search where loadings are selected from Hausman weights {-1,0,1} at random at each step for the same number of iterations as the simple component algorithm.

2.5.2 The Choice of Penalty Parameter

The variable selection algorithm proposed by Cadima and Jolliffe (2001) was used to extract a twelve variable subset from the deodorant sensory data introduced in Section 1.1.3. This subset best correlated with the full set of principal components. The variables selected are shown in Table 2.6 and these are the ones which will be used to assess the simple component algorithm and choice of the penalty parameter. Simple components using an orthogonal loss (2.5) were extracted from the 12 variable subset using the correlation matrix. Figures 2.1, 2.2 show how the simple component algorithm performs over a range of tuple sizes from 1 to 11 for the penalty parameter set to 0.1, 1 and 10. The full set of 12 components are considered. Notice that the CSV is less than one across the range of k. In all cases the systems are correlated. The loss of orthogonality for the systems under the three chosen penalty parameters indicates that even for a penalty parameter of 10 there is some loss of orthogonality across the range of k and is 1.9 at its lowest for the full set of 12 components. Clearly for a correlation matrix this suggests that the penalty parameter should be larger than 1.0. In Figure 2.1 the CSV is slightly better at lower k for a penalty of 10, however

Overall opinion of effectiveness
Dried quickly
Rollball glided over skin
Fragrance lasted long enough for me
Marked clothes
How sticky whilst applying
Gave me day long protection against wetness
Pack did not become messy
Easy to apply the right amount
How cold whilst applying
How easy to wash off skin
How often applied rollon

Table 2.6: The 12 selected variables

this can be explained by the slight loss of orthogonality shown in 2.2. Also even for a penalty of 10 the CSV $< Opt_1$ indicating that the component scores are correlated. For k=2 or 3 the values of CSV, Opt_1 and the loss of orthogonality are not very different for those for larger values of k. Figures 2.3, 2.4[1] show similar plots for penalties 10, 100 and 1000 for k 1 to 11. The penalty parameter of 1000 shows the smallest loss of orthogonality. If the loss is small for $k < 4$ then the algorithm may be useful for large data sets.

When using a penalty based on the squared correlation between components, equation (2.6), the picture is different. For this problem τ should be greater than one, however once τ is of the of order of 100 or more the performance drops. Figures 2.5 and 2.6, show the performance for τ of 0.1, 1 and 10. Figures 2.7 and 2.8 show the CSV and Opt_1 for values of τ of 10, 100 and 1000. In Figure 2.7 the CSV is greater and Opt_1 closer to unity when τ is 10 than when it is 100 or 1000 and this is the case for all values of k. An explanation for this may be that because the algorithm is sequential, if a large τ finds a very good solution for the first components, the second or third for instance, then this may prevent better solutions being found later and thus reduce the performance of the complete set.

Variability

Figure 2.9 shows the variability in the CSV and Opt_1 performance measures for an orthogonal penalty and Figure 2.10 for correlation penalty. The variability was calculated for twenty runs for each tuple size k. and then calculating the mean and standard

[1]It might be expected that the loss of orthogonality would decrease monotonically with increasing k, but this is not the case in Figure 2.4. However, as the greedy algorithm is sequential and finds the best local solution at each step, monotonicity cannot be guaranteed. For example, very good solutions found early can prevent a better global solution later.

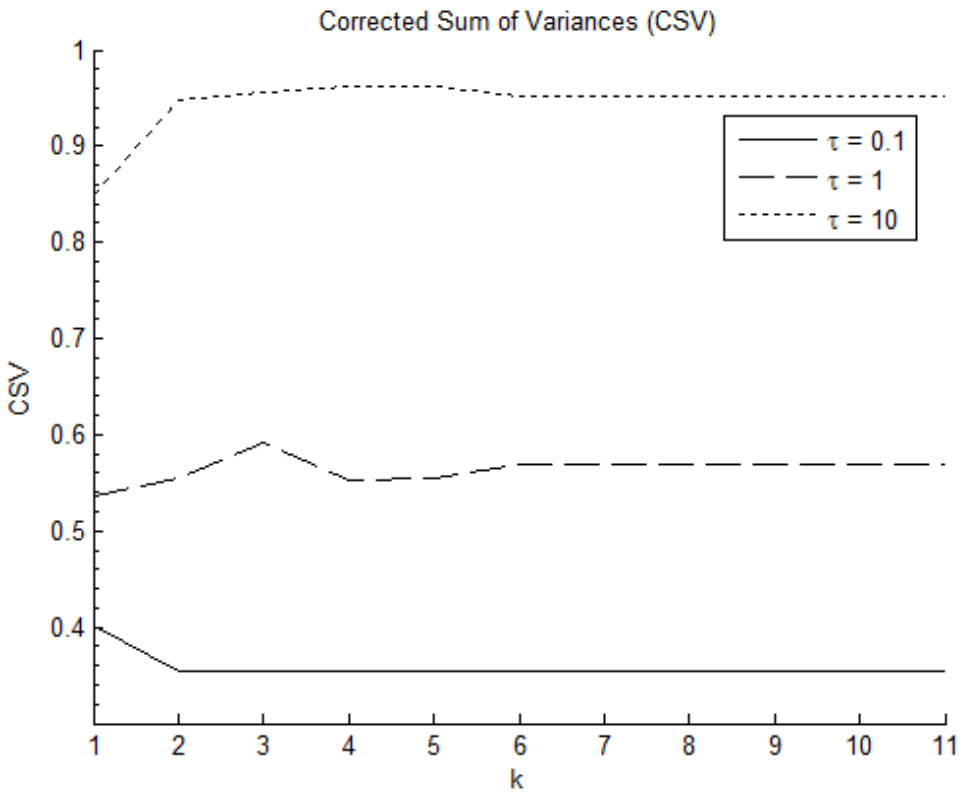


Figure 2.1: The CSV for simple components using a squared orthogonal penalty. These are plotted for different orders of the penalty parameter (τ).

deviation of the CSV and Opt_1 measures. Figures 2.9, 2.10 show the variability in the CSV and Opt_1 over the lower k range, 1 to 5. The plots are based on ten repeats, for $\tau = 1000$ for an orthogonal penalty and $\tau = 10$ for a squared correlation penalty. The algorithm was seeded with random simple vectors. Variability, even using lower k values, 2 or 3 is low.

One of the intended applications of simple components in this thesis is the application to large data sets. The major computational overhead is the choice of k. If the performance of the algorithm is good with small k then this makes large problems tractable. To this end reducing k, but increasing the the number of restarts within the algorithm may provide a speed up. For each component the best solution found after a complete step can be fed back in as the new starting vector.

Meta-heuristic Choice for τ

For both the orthogonal and correlation penalty, extreme values of τ may not produce the best solutions. At the lower extreme where τ is close to zero, a solution will be obtained where all the components are identical, and the loss of orthogonality is the

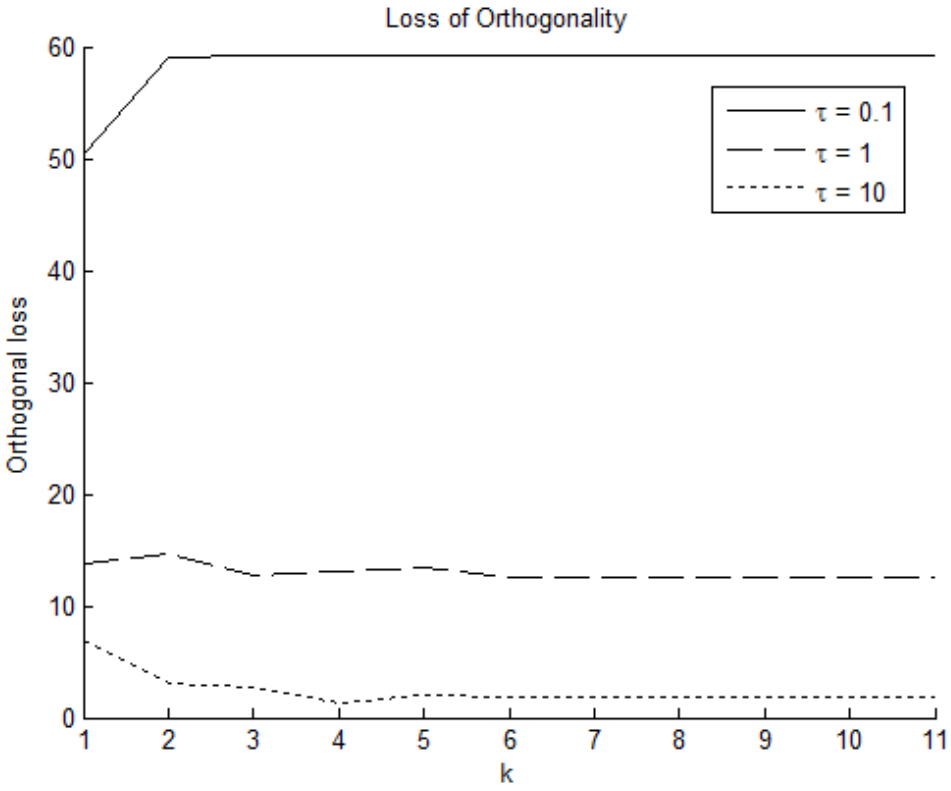


Figure 2.2: The loss of orthogonality for simple components using a squared orthogonal penalty. These are plotted for different orders of the penalty parameter (τ).

number of variables, p. At the other extreme, when the loss function is driven by the penalty, directions of high variability are missed when the loss of orthogonality is very small but acceptable. So low or very high penalty values seem a bad choice.

There is no way to determine the 'best' value for the penalty parameter a priori and the choice will be problem dependent. If for instance a covariance matrix is used then the variances may be large, however the loss is always bounded [2]

$$0 \leq \text{ loss } \leq p - 1.$$

The variance is partitioned by the eigenvalues of the covariance matrix and these can provide some guidance as to the magnitude of the penalty parameter.

One strategy is to use the eigenvalue corresponding to the last found component, and use this as the current penalty parameter. As the variation in the data can be

[2] At each step the current solution is assessed for its squared orthogonal loss against the previously found loading vectors. The largest this can be is $p-1$ as in the case of the final pth loading vector. In the general case the maximum loss is the sum of the squared loss between all pairs of loading vectors and so for the full set is $p(p-1)/2$. In Figure 2.6 for example, where $p = 12$, the maximum orthogonal loss of the system is 66.

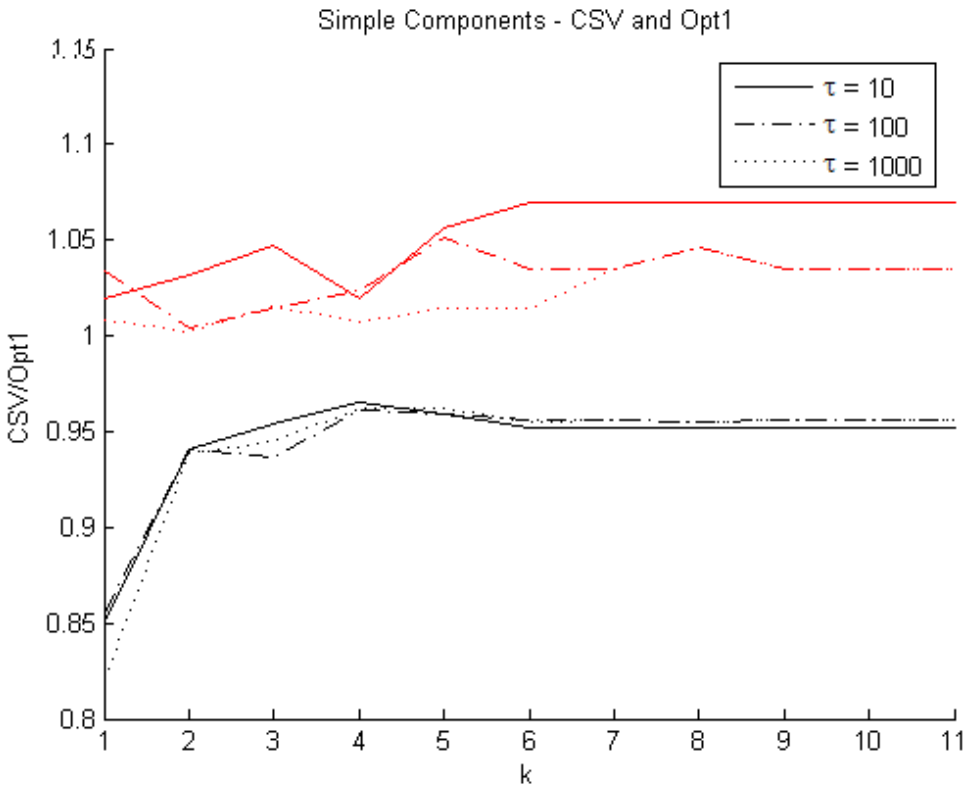


Figure 2.3: The CSV and Opt_1 is plotted for the twelve variable deodorant data for large penalties. The lower lines are the CSV values and the upper the corresponding Opt_1 values. The difference between Opt_1 and the CSV values is a measure of the correlation of the component system.

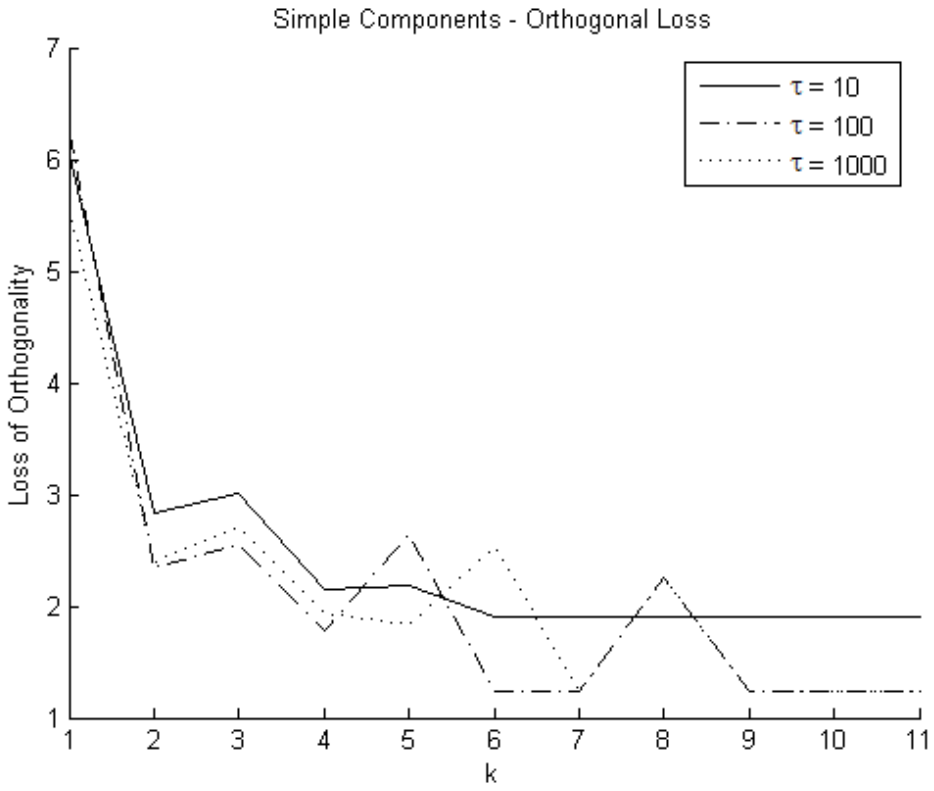


Figure 2.4: The orthogonal loss plotted for the squared orthogonal penalty when $\tau \in \{10, 100, 1000\}$

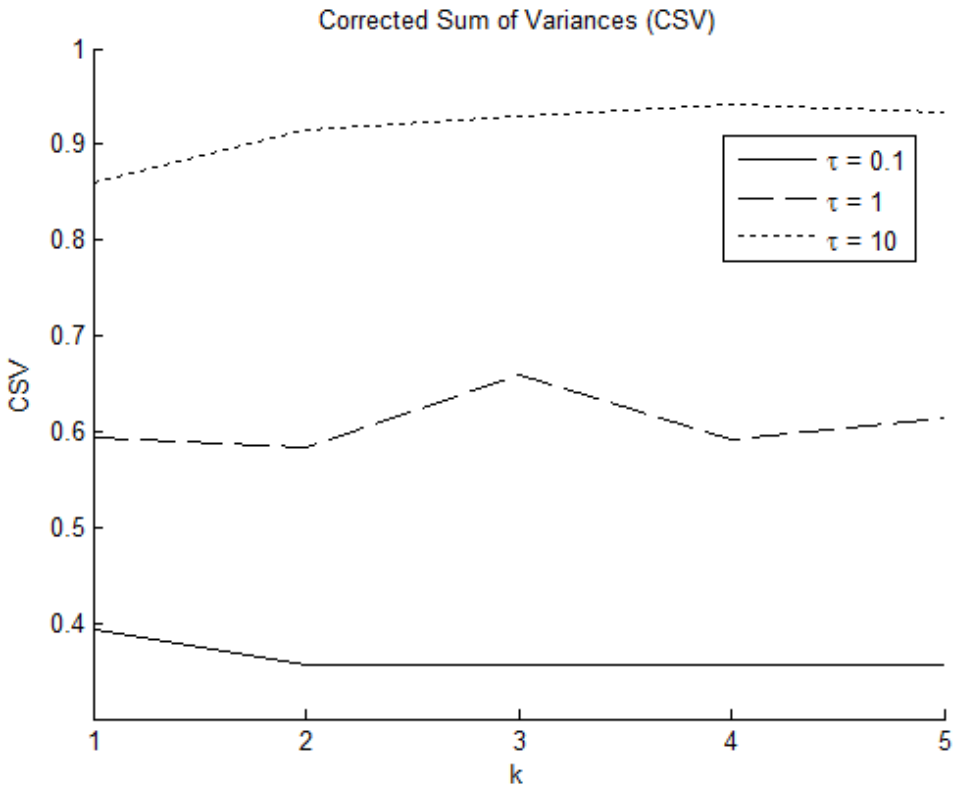


Figure 2.5: The CSV for values of $\tau \in \{0.1, 1, 10\}$ when the penalty is based on a squared correlation.

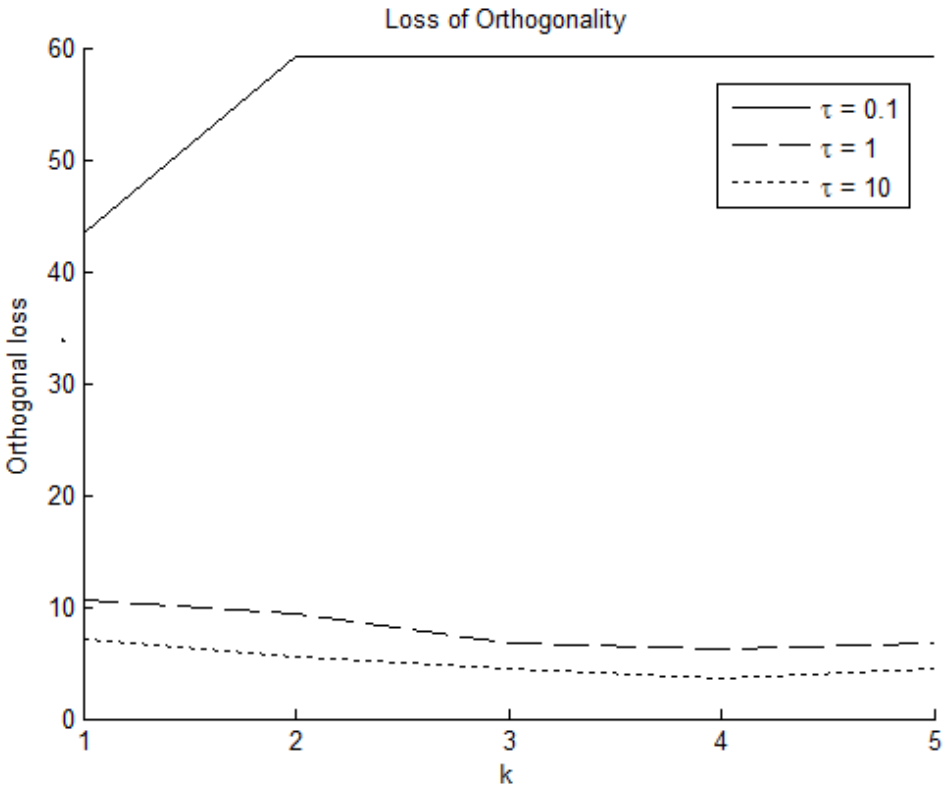


Figure 2.6: The loss of orthogonality for a penalty based on a squared correlation, and $\tau \in \{0.1, 1, 10\}$.

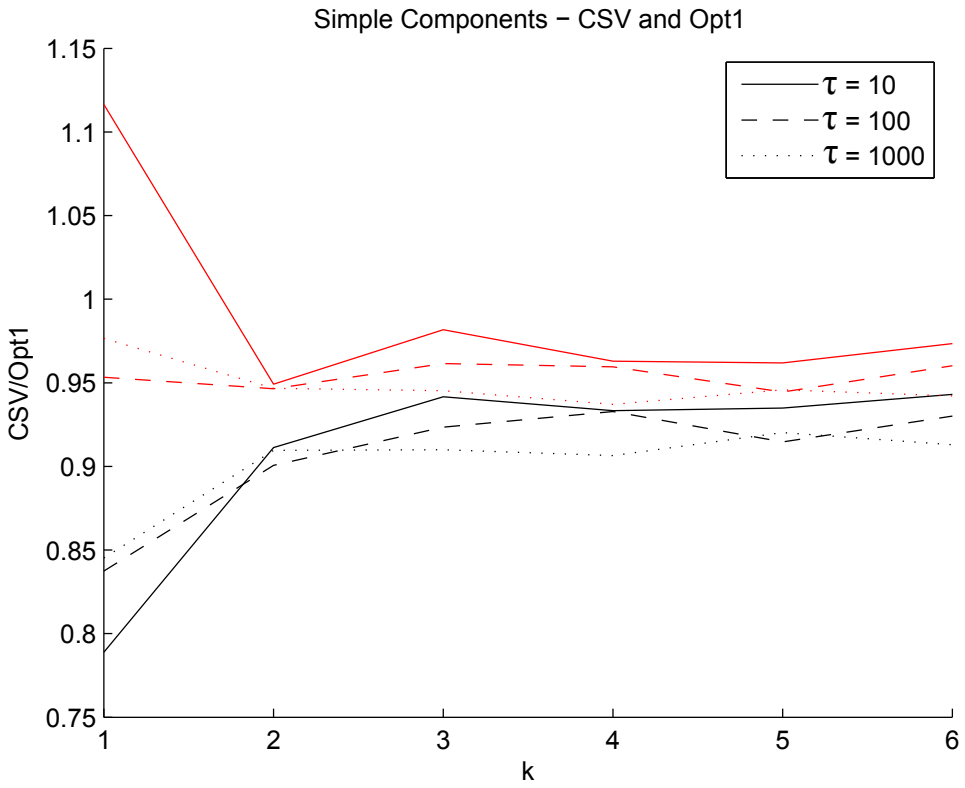


Figure 2.7: The CSV (lower plots) and Opt_1 (upper plots) based on a squared correlation penalty, for high values of the penalty parameter, τ.

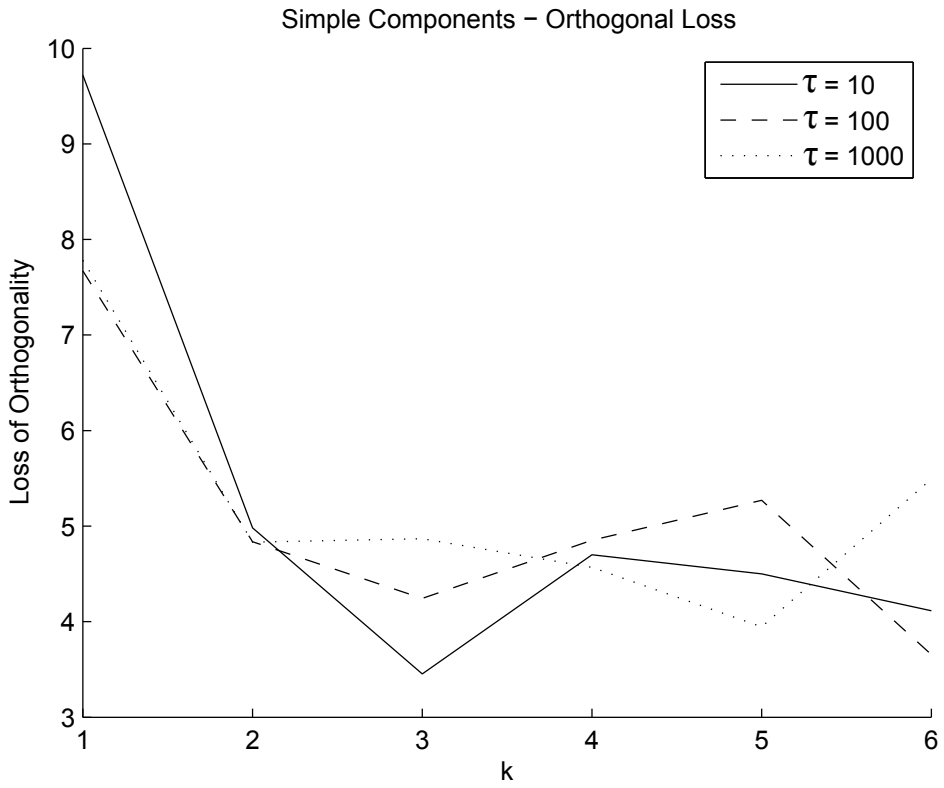


Figure 2.8: The loss of orthogonality for a penalty based on a squared correlation, for larger penalty parameter values, τ.

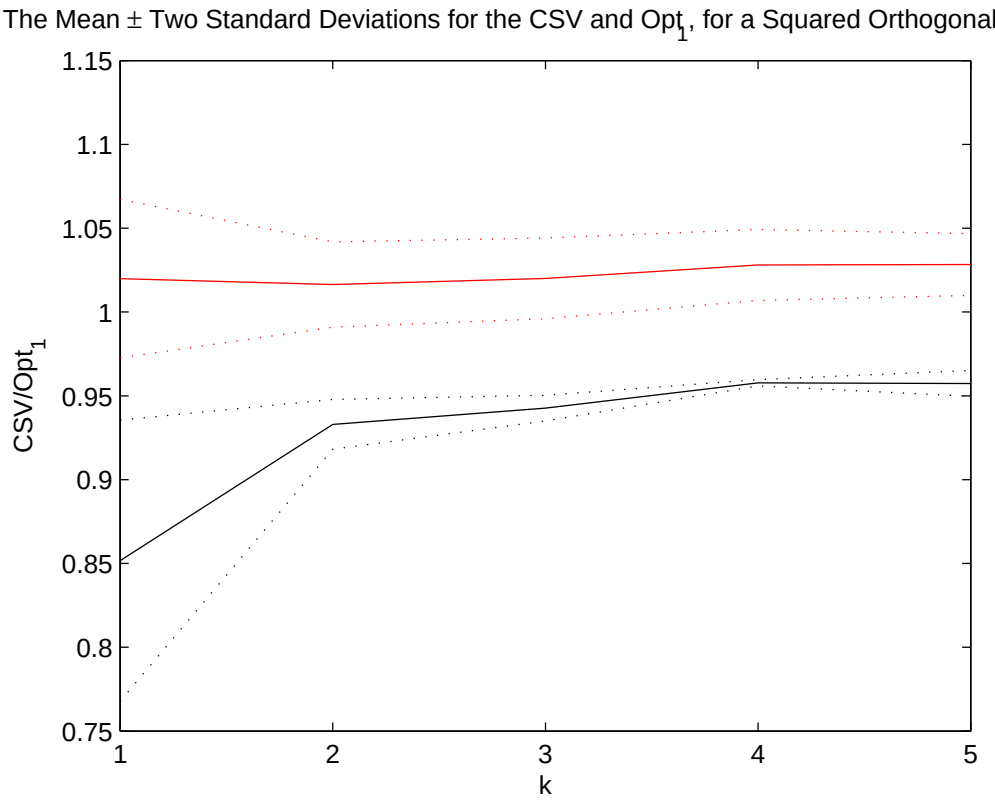


Figure 2.9: The reproducibility in the CSV and Opt_1 values for simple components using a squared orthogonal penalty.

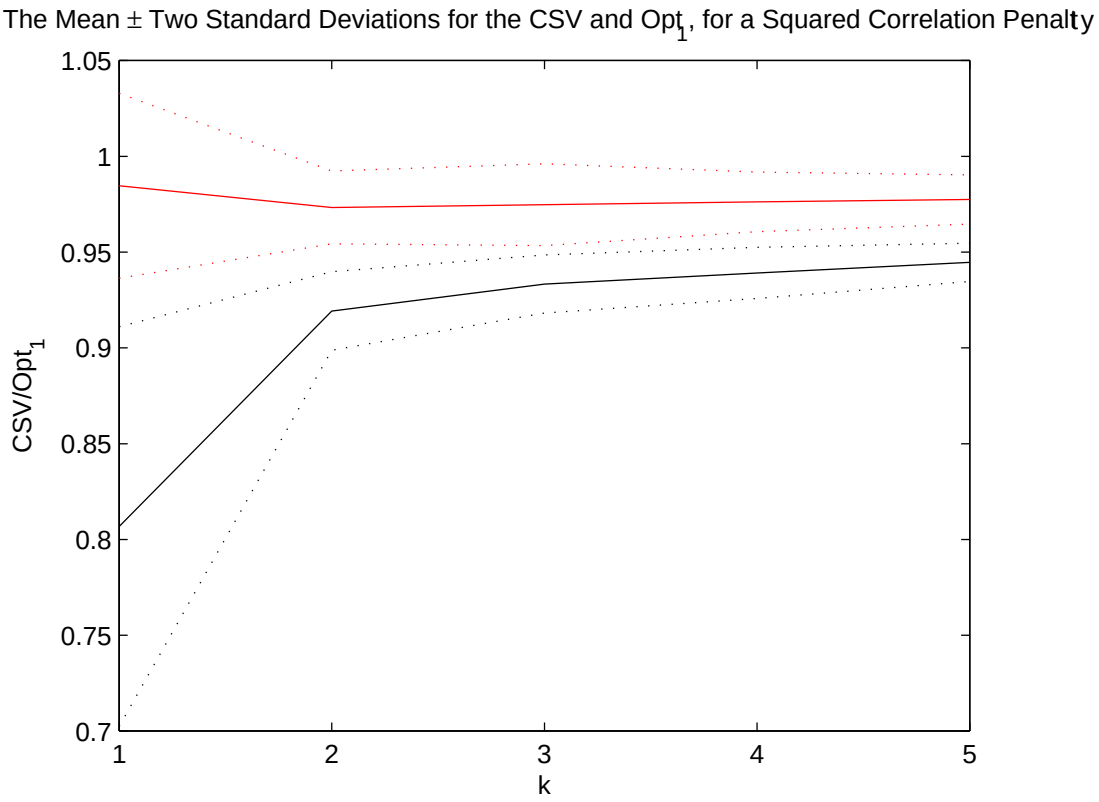


Figure 2.10: The reproducibility in the CSV and Opt_1 for simple components using a squared correlation penalty.

partitioned exactly using the eigenvalues these may provide a heuristic approach to determine some bounds on τ. A simple idea is to use the total variation in the data as a penalty or the largest eigenvalue. If a component becomes close to the first principal component then this incurs a penalty of λ_1 and the objective function is close to zero. For a given component $\mathbf{w}_k$ its variance will be bounded by the eigenvalues λ_{k-1} and λ_{k+1}. If an acceptable loss of orthogonality is defined as δ, then one strategy is to choose τ_k to be

$$\tau_k = \frac{(\lambda_{k-1} - \lambda_{k+1})}{\delta}.$$

The problem of choosing τ now becomes that of choosing δ. The closeness of two unit vectors is measured by the cosine of the angle θ between them,

$$\mathbf{w}_i' \mathbf{w}_j = \cos\theta.$$

Defining δ in terms of this angle,

$$\mathbf{w}_i' \mathbf{w}_j = \cos\left(\frac{\pi}{2} - \delta\right) = \sin\delta.$$

If an acceptable loss is chosen to be five degrees then δ is 0.0872 radians and the objective function becomes

$$\max_{w_k} \frac{\mathbf{w}_k' \mathbf{S} \mathbf{w}_k}{\mathbf{w}_k' \mathbf{w}_k} - 11.5 \times (\lambda_{k-1} - \lambda_{k+1}) \times \text{loss}$$

Another approach is to recognise that early on there are more degrees of freedom available in the data in which to find components of high variability that are orthogonal. It may make sense to weight the loss less severely than later when the amount of variation available and the degrees of freedom are less. A strategy is to find $\mathbf{w}_1$ from

$$\max_{w_1} \frac{\mathbf{w_1}' \mathbf{S} \mathbf{w_1}}{\mathbf{w_1}' \mathbf{w_1}} < \lambda_1, \tag{2.11}$$

which must be less than the eigenvalue of $\mathbf{u}_1$. Now use λ_1 as the penalty,

$$\max_{w_2} \frac{\mathbf{w_2}' \mathbf{S} \mathbf{w_2}}{\mathbf{w_2}' \mathbf{w_2}} - \lambda_1 \times \text{loss}. \tag{2.12}$$

The process continues by finding $\mathbf{w}_3$ subject to $\tau = \lambda_1 + \lambda_2$ and so on. So for the final component,

$$\max_{w_p} \frac{\mathbf{w_p}' \mathbf{S} \mathbf{w_p}}{\mathbf{w_p}' \mathbf{w_p}} - \sum_{i=1}^{p-1} \lambda_i \times \text{loss} \tag{2.13}$$

Six strategies are compared, which are listed in Table 2.7. Table 2.8 show the performance values obtained for each of the strategies.

All the solutions for both the full set of twelve components, and five components, indicate that setting τ_k to the current eigenvalue is a poor choice (strategy 5). In all

1. The total variation $\tau = \sum_{i=1}^{p} \lambda_i$
2. The largest eigenvalue $\tau = \lambda_1$
3. $\tau_k = (\lambda_{k-1} - \lambda_{k+1}) / \delta$, where δ is the angle in radians. In the case of an orthogonal penalty, δ refers to the angle between the component axes. For a correlation penalty, this is the loss of orthogonality between the component scores. In this case an orthogonal loss of five degrees between component axes was chosen, expressed in radians.
4. $\tau_k = \sum_{i=1}^{k-1} \lambda_i$
5. $\tau_k = \lambda_k$
6. The CSV used as an optimization criterion, i.e. replacing equations (2.5) or (2.6). No penalty parameter is required as the CSV is unique and bounded.

Table 2.7: Strategies to choose suitable values of the penalty parameter τ

Approach	Penalty Type	k	q	CSV	Opt1	System Correlation (Opt_1 - CSV)	Orthogonal loss
1	Orthogonal	2	12	0.93	0.97	0.11	3.12
2				0.91	1.26	0.26	4.56
3				0.94	0.95	0.09	2.55
4				0.93	0.98	0.15	3.98
5				0.39	2.04	2.32	34.22
6				0.96	1.01	0.05	3.28
1	Correlation	2	12	0.91	0.97	0.05	5.45
2				0.75	1.26	0.51	5.60
3				0.91	0.95	0.04	4.85
4				0.89	0.98	0.08	5.73
5				0.56	2.04	1.48	13.10
1	Orthogonal	2	5	0.86	0.92	0.02	0.00
2				0.94	0.92	0.06	0.55
3				0.90	0.89	0.01	0.00
4				0.91	0.94	0.06	0.45
5				0.65	1.32	1.11	5.12
6				0.95	0.97	0.01	0.60
1	Correlation	2	5	0.92	0.92	0.00	0.20
2				0.92	0.92	0.00	0.27
3				0.88	0.89	0.00	0.59
4				0.94	0.94	0.00	0.18
5				0.79	1.32	0.53	1.53

Table 2.8: Table showing the performance of each of the penalty strategies listed in 2.7. Solutions were found for five and twelve components (q), for a tuple size of two (k).

cases there is a relatively large loss of orthogonality compared to the other strategies and the CSV is low. This is more marked for an orthogonal penalty than a correlation penalty. Strategies, 1, 3, 4 and 6 perform well, but 2 does not perform so well on the full set of components, particularly when a correlation penalty is employed. In fact the system correlation is higher (0.51) than for the corresponding orthogonal penalty (0.26). There are differences between the use of an orthogonal penalty and a correlation penalty. Strategy 3 performs better with an orthogonal penalty. Using the CSV (strategy 6) produces good results across the board, as expected it focusses on reducing the system correlation at the expense of orthogonality, however, this loss is still comparable with other strategies.

Benchmarks

Figures 2.11, 2.12, show the performance of simple components against random search and full enumeration. Five components of the twelve possible were extracted. A penalty of twelve was chosen, which is the total variation using the correlation matrix. Enumeration achieves a CSV of 0.95, a Opt_1 of 1.07, and a loss of orthogonality of 1.90. This indicates that the component scores are slightly correlated. For $k = 3$ or 4, simple components is performing close to that of sequential enumeration. The random search performs significantly worse.

The component loadings for each algorithm are shown in Table 2.9. Both simple components and enumeration find the same first component, however the second differ slightly in that *how easy to wash off skin* and *how often applied rollon* are reversed in sign. The enumerated solution explains more variation and an interpretation could be a contrast between efficacy and product properties, where *how often applied rollon* is correlated with efficacy and *how easy to wash off skin* with product properties. E7 is extremely sparse and is a simple average of how easy to wash off the product and how often it is applied. None of the simple components are this sparse, but C3 is a simple contrast between the deodorant's properties on the skin versus how often it is applied and some aspects of its efficacy i.e. wetness protection and fragrance.

Reconstruction Error

Figures 2.13 shows the reconstruction error, as described in Section 2.5.1, for the full set of sensory data introduced in Section 1.1.3, consisting of forty nine variables. The unexplained variance is calculated as each component is added to the component set. PCA is optimal and has the smallest reconstruction error possible. Notice that for simple components where k is two or three, the unexplained variance is favourable. Figure 2.14 shows the unexplained variance for the first ten components. The variance

Simple Components (k=3)	C1	C2	C3	C4	C5	C6	C7
overall opinion - effective	-1	-1	0	0	-1	-1	0
dried quickly	-1	0	1	1	0	0	1
rollball glided over skin	-1	1	0	-1	-1	0	1
fragrance lasted long enough for me	-1	-1	-1	1	1	-1	0
marked clothes	1	0	0	-1	0	-1	1
how sticky whilst applying	-1	0	1	1	-1	0	0
gave me day-long protection- wetness	-1	-1	-1	-1	-1	1	0
pack did not become messy	-1	1	0	-1	0	0	-1
easy to apply the right amount	-1	1	0	-1	1	-1	0
how cold whilst applying	-1	0	0	0	1	1	1
how easy to wash off skin	0	-1	1	-1	0	-1	0
how often applied rollon	0	1	-1	1	-1	-1	1
Variance explained	**3.73**	**1.14**	**1.14**	**0.87**	**0.84**	**0.80**	**0.76**
CSV	**0.95**	Opt_1	**1.02**	Opt_2	**0.07**	**Orth loss**	**2.90**

Enumeration	E1	E2	E3	E4	E5	E6	E7
overall opinion - effective	1	1	1	0	1	1	0
dried quickly	1	0	0	1	-1	1	0
rollball glided over skin	1	-1	1	0	1	-1	0
fragrance lasted long enough for me	1	1	-1	0	1	0	0
marked clothes	-1	0	1	0	1	1	0
how sticky whilst applying	1	0	0	1	-1	1	0
gave me day-long protection- wetness	1	1	-1	0	1	0	0
pack did not become messy	1	-1	-1	-1	0	1	0
easy to apply the right amount	1	-1	-1	-1	0	1	0
how cold whilst applying	1	0	-1	0	-1	-1	0
how easy to wash off skin	0	-1	-1	1	1	0	1
how often applied rollon	0	1	1	-1	-1	0	1
Variance explained	**3.73**	**1.18**	**1.09**	**1.07**	**1.03**	**0.92**	**0.89**
CSV	**0.95**	Opt_1	**1.07**	Opt_2	**0.12**	**Orth loss**	**1.90**

Random Search (k=3)	R1	R2	R3	R4	R5	R6	R7
overall opinion - effective	-1	0	1	0	-1	1	-1
dried quickly	-1	0	0	0	0	0	1
rollball glided over skin	0	1	1	-1	1	1	1
fragrance lasted long enough for me	-1	-1	1	1	1	0	0
marked clothes	1	-1	0	-1	0	1	1
how sticky whilst applying	-1	0	0	-1	-1	-1	1
gave me day-long protection- wetness	-1	0	0	1	-1	1	1
pack did not become messy	0	1	1	0	-1	-1	1
easy to apply the right amount	-1	1	-1	0	1	0	0
how cold whilst applying	-1	0	-1	-1	-1	1	0
how easy to wash off skin	0	0	1	-1	0	0	-1
how often applied rollon	1	1	0	1	-1	1	0
Variance explained	**3.03**	**1.43**	**0.97**	**0.97**	**0.93**	**0.82**	**0.80**
CSV	**0.87**	Opt_1	**1.04**	Opt_2	**0.17**	**Orth loss**	**4.52**

Table 2.9: The loadings for the benchmark example, showing simple components, sequential enumeration and random search.

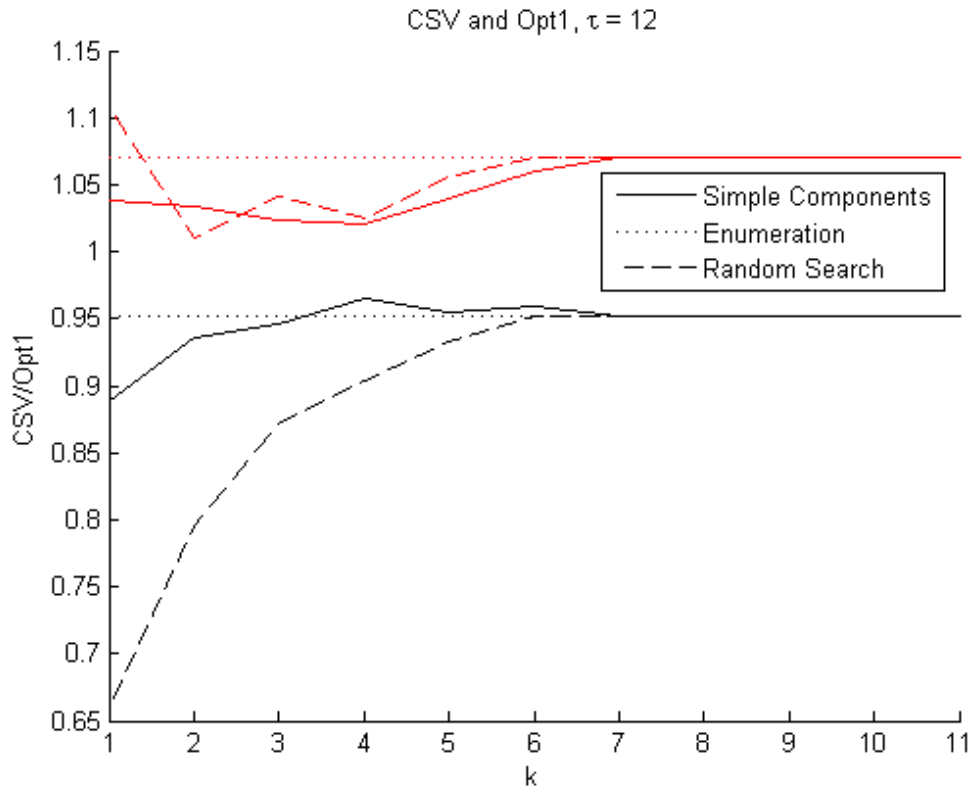


Figure 2.11: The comparison of the simple component algorithm against the benchmarks (enumeration and random search). CSV (lower plots) and Opt_1 (upper plots).

explained and cumulative variance explained compared to PCA is shown in Figures 2.15 and 2.16. From the cumulative percentage of variance explained, 80 % of the variation in the data is explained by twenty principal components, and thirty simple components.

2.5.3 Adaptations

The simple component search finds simple axes which maximize the observed variance in a data set. Orthogonal axes or uncorrelated scores are preferred by penalizing with the penalty functions described in equations (2.5) and (2.6). However, these rely on the choice of a penalty parameter to provide near orthogonal axes or lower correlated scores. The CSV criterion is generally applicable and unique for any $\mathbf{A}_q$ and assesses the loss of orthogonality and the increase in correlation compared to PCA. It can be used as an optimization criterion, equation (2.10). This criterion is available in the R package by Rousson and Maechler (2009), which finds simplified components (Rousson and Gasser, 2004).

An alternative approach is to use a criterion based on a modification of the CSV. The objective in equation (2.14) could be used to favour solutions that have orthogonal

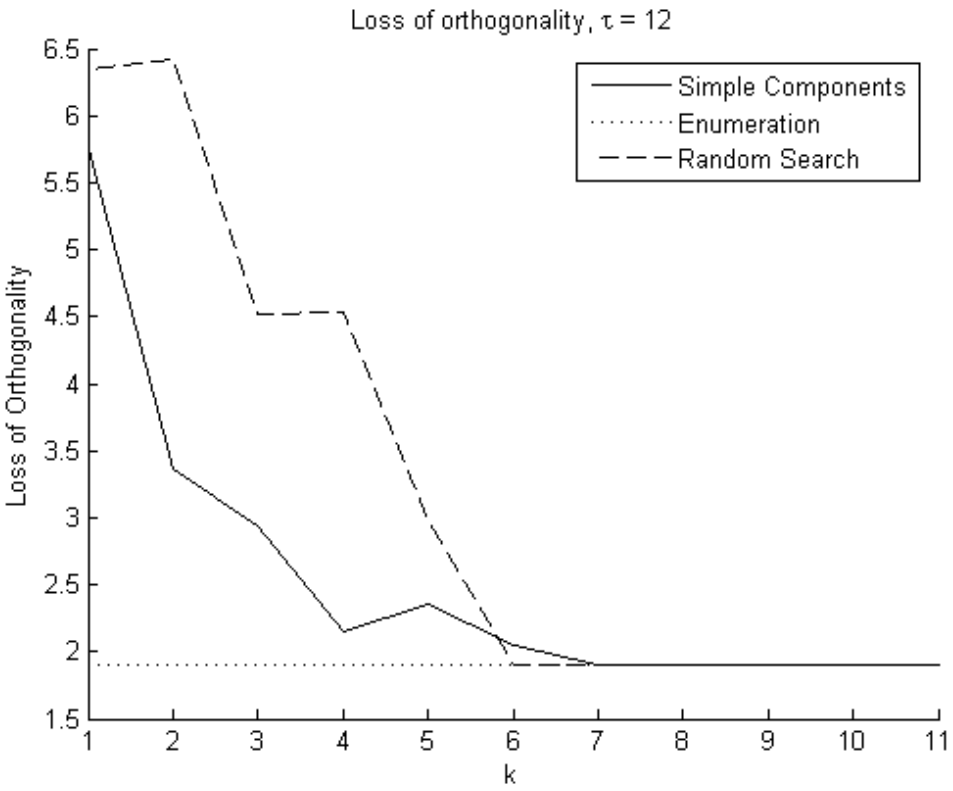


Figure 2.12: The comparison of the simple component algorithm against the benchmarks (enumeration and random search). Loss of orthogonality.

weight vectors but with highly correlated scores. Opt_2 will select components with correlated scores, while Opt_1 will be one if the system is orthogonal, otherwise it can be greater than or smaller than one. So $(1 - Opt_1)^2$ can be used to favour solutions where Opt_1 tends to one.

$$\max_a \ \mathrm{Opt}_2 - \tau \times (1 - Opt_1)^2 \tag{2.14}$$

PCA is a form of projection pursuit where the criterion is to reduce the dimensionality of a set of multivariate data by finding directions of maximum variance. However, other criterion are possible. For example an information theoretic measure such as entropy. The simple component framework is combinatorial and can be adapted to use any optimization criterion. For instance, including a simplicity function such as varimax, or the L_1 norm constraint on the weight vector to improve robustness.

2.6 Data Examples

The simple component analysis algorithm of Rousson and Gasser (2004) is implemented as a package for the R project for statistical computing Rousson and Maechler (2009).

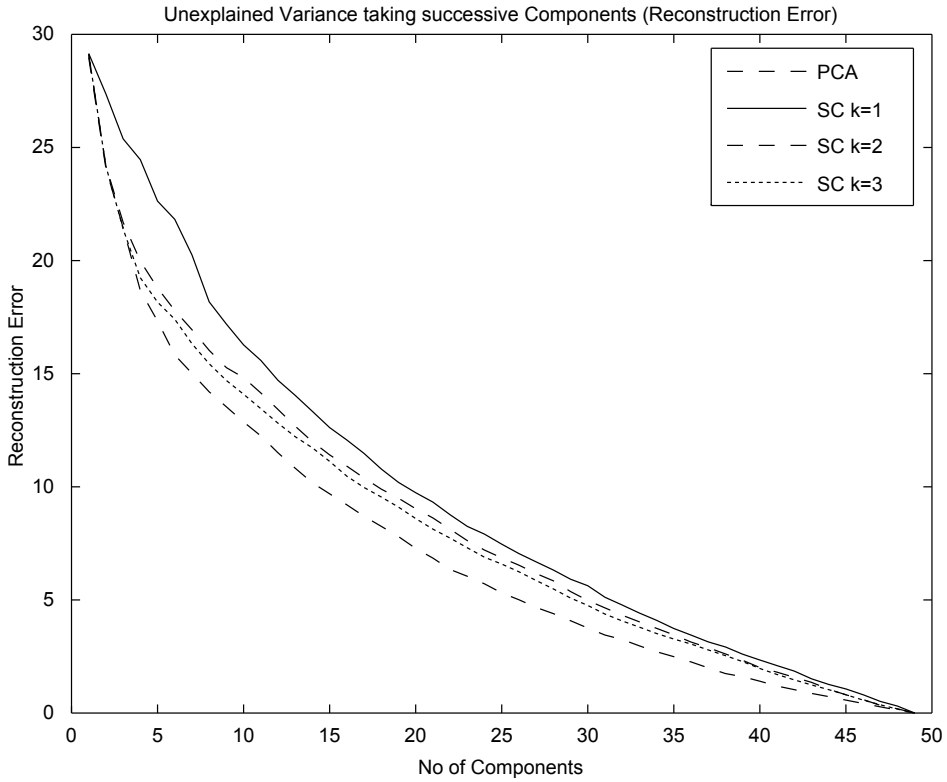


Figure 2.13: Reconstruction Error for PCA (lowest line) and SC over the full set of forty nine components.

This provides correlation matrices for small data sets which have been analysed in the literature. The pitprops data arose from a study on the strength of pitprops cut from timber. The correlation matrix is obtained from thirteen variables which have the following meaning, Table 2.10,

Code	**Description**
TOPDIAM	Top diameter of the prop in inches
LENGTH	Length of the prop in inches
MOIST	Moisture content of the prop, expressed as a percentage of the dry weight
TESTSG	Specific gravity of the timber at the time of the test
OVENSG	Oven-dry specific gravity of the timber
RINGTOP	Number of annual rings at the top of the prop
RINGBUT	Number of annual rings at the base of the prop
BOWMAX	Maximum bow in inches
BOWDIST	Distance of the point of maximum bow from the top of the prop in inches
WHORLS	Number of knot whorls
CLEAR	Length of clear prop from the top of the prop in inches
KNOTS	Average number of knots per whorl
DIAKNOT	Average diameter of the knots in inches

Table 2.10: Variable labels and descriptions for the pitprop data

The thirteen extracted simple components and the performance measures are shown

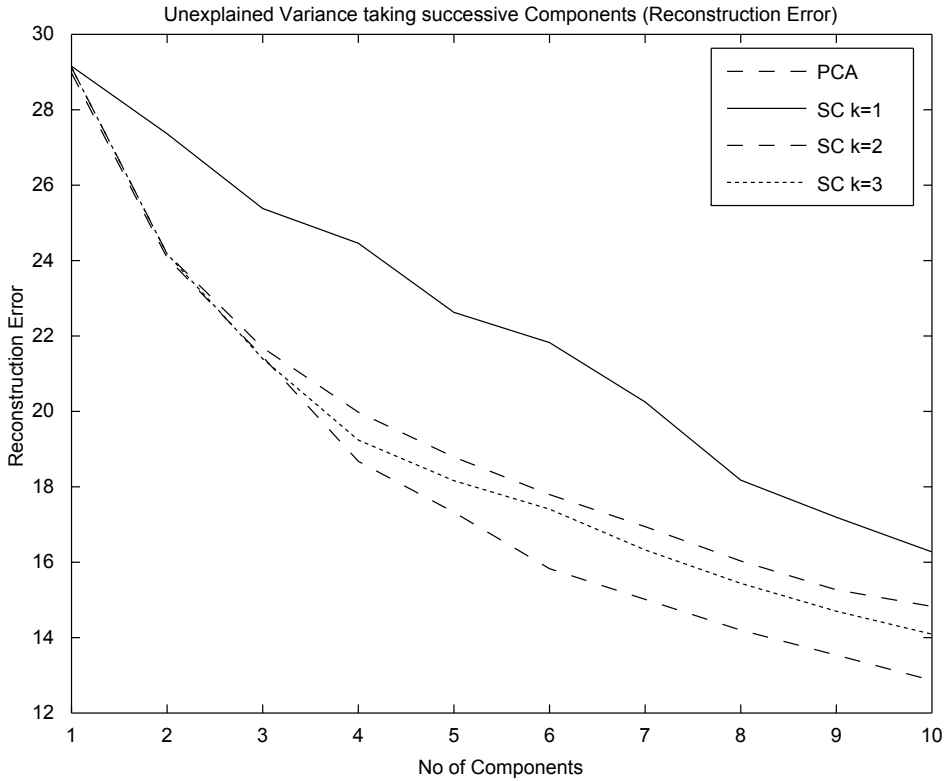


Figure 2.14: Reconstruction Error for PCA (lowest line) and SC showing the first ten components.

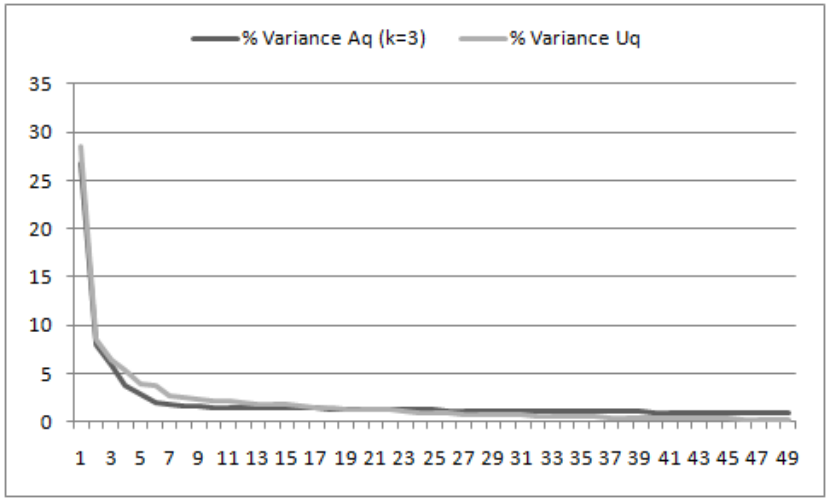


Figure 2.15: Percentage variance explained for PCA ($\mathbf{U}_q$) and simple components ($\mathbf{A}$)

in Table 2.11. Notice that in all cases the first two components are identical. The simple components set, using $k = 2$ is slightly more correlated than the enumerated set, otherwise the performance is comparable. Opt_2 is a measure of the correlation present within the component system. With a tuple size of $k = 2$, C1 and C5 split the variables into a weighted average and a contrast between the remaining variables.

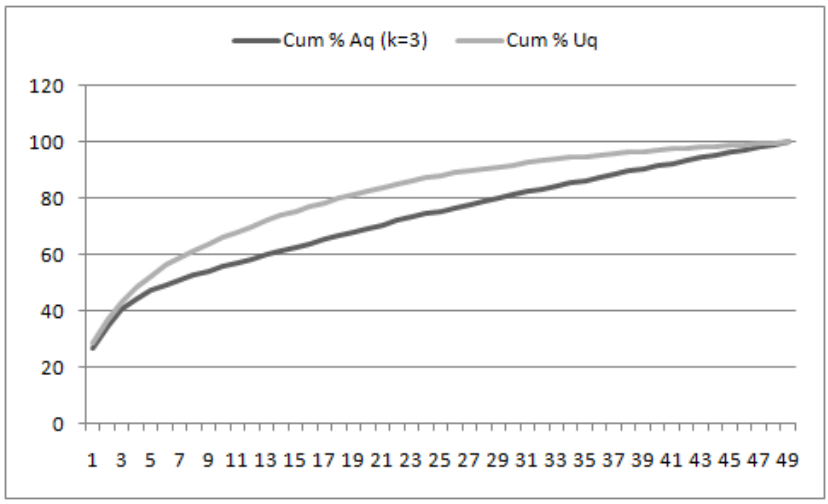


Figure 2.16: Cumulative percentage variance explained for PCA ($\mathbf{U}_q$) and SC ($\mathbf{A}_q$), τ = 1000.

MOIST, TESTSG, OVENSG and DIAKNOT is contrasted with KNOTS and CLEAR.

Jolliffe (2002), page 286, compares SCoT, SCoTLASS and the components obtained by Vines (2000), for the first and fourth components. Interestingly, SCotLASS produces a structure for the first component similar to C1, where there is a small non-informative weight of 0.05 for RINGTOP, C1 includes this variable. The first principal component loadings could be interpreted as a weighted average excluding variables 5, 11 ,12 and 13. However, this is a subjective interpretation. For the fourth component, SCoTLASS has the simplest interpretation in Jolliffe, but is hampered by a number of small difficult to interpret weights. Vines's simple component is not at all simple and in fact is similar to interpreting the principal component, as the integer loadings are large and of varying magnitude making a number of them non-informative. However, the new algorithm produces a very clear contrast. C7 is interesting as it is sparse, in fact with this simple component algorithm, interesting contrasts are extracted for low variance components, which are often simple to interpret. One difficulty is that solutions obtained for $k = 2$ and $k = 4$ do differ from each other and from the sequential enumeration. It is not clear what is the best combination of k and the number of algorithm restarts. As mentioned with a sequential greedy approach, good solutions early on may prevent better solutions being found for the whole set. Using the greedy search in a simultaneous manner as described in Section 2.4 may be tractable for problems of this size.

2.7 Re-Analysis of the Sensory Panel Data

Table 2.12 show the first five unscaled simple component loadings for the sensory panel data introduced in Section 1.1.3. The components were obtained using a tuple size of

Simple components	k=2						
Code	C1	C2	C3	C4	C5	C6	C7
TOPDIAM	-1	-1	0	0	0	0	0
LENGTH	-1	-1	0	0	0	0	0
MOIST	0	-1	1	1	1	-1	0
TESTSG	0	-1	1	1	1	0	0
OVENSG	0	1	1	0	1	1	1
RINGTOP	-1	0	1	0	0	1	-1
RINGBUT	-1	1	1	0	0	0	-1
BOWMAX	-1	0	-1	1	0	0	1
BOWDIST	-1	0	0	0	0	0	1
WHORLS	-1	1	0	0	0	-1	0
CLEAR	0	-1	0	1	-1	1	0
KNOTS	0	-1	1	-1	-1	0	1
DIAKNOT	0	-1	-1	-1	1	1	0
Variance explained	**3.92**	**2.15**	**1.71**	**1.17**	**0.72**	**0.45**	**0.66**
CSV	**0.85**	Opt_1	**1.03**	Opt_2	**0.18**	**Orth loss**	**2.9**
Code	k=4						
TOPDIAM	-1	-1	0	0	0	0	0
LENGTH	-1	-1	0	0	0	0	0
MOIST	0	-1	1	1	1	0	0
TESTSG	0	-1	1	1	1	-1	0
OVENSG	0	1	1	1	-1	-1	1
RINGTOP	-1	0	1	-1	-1	-1	0
RINGBUT	-1	1	1	0	1	1	0
BOWMAX	-1	0	-1	1	-1	-1	-1
BOWDIST	-1	0	-1	1	0	1	1
WHORLS	-1	1	0	0	1	0	0
CLEAR	0	-1	0	1	-1	1	0
KNOTS	0	-1	1	-1	-1	1	0
DIAKNOT	0	-1	-1	-1	1	-1	1
Variance explained	**3.92**	**2.15**	**1.71**	**1.25**	**0.81**	**0.67**	**0.70**
CSV	**0.88**	Opt_1	**1.05**	Opt_2	**0.17**	**Orth loss**	**3.34**
Enumeration	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**
TOPDIAM	-1	-1	0	1	0	0	1
LENGTH	-1	-1	0	0	0	0	1
MOIST	0	-1	1	1	0	0	-1
TESTSG	0	-1	1	1	1	0	0
OVENSG	0	1	1	0	1	-1	0
RINGTOP	-1	0	1	-1	0	0	1
RINGBUT	-1	1	1	0	0	0	0
BOWMAX	-1	0	-1	0	1	0	-1
BOWDIST	-1	0	0	0	0	0	-1
WHORLS	-1	1	0	1	-1	0	-1
CLEAR	0	-1	0	-1	1	0	-1
KNOTS	0	-1	1	-1	-1	0	-1
DIAKNOT	0	-1	-1	0	-1	-1	0
Variance explained	**3.92**	**2.15**	**1.90**	**1.26**	**1.05**	**0.79**	**0.52**
CSV	**0.90**	Opt_1	**1.09**	Opt_2	**0.19**	**Orth Loss**	**3.10**

Table 2.11: Simple components and enumeration of the pitprops strength data. An orthogonal penalty was used

three, and the adaptive penalty (4) of Table 2.8. The variables have been clustered into groups in a similar way to the introduction (see, Section 1.1.3). SC1 is a contrast between group 3 and the other variables. Group 2 could be further split. The last four variables in that group are mainly zero across the five components. This indicates that they are not important to the respondents or did not differ much between products in the study, and this is quite clear from the component loadings. For instance trapping of underarm hair may differentiate between the products tested, but are only on the fifth component. SC2 can be interpreted as contrasting efficacy and fragrance with greasy, sticky and how easy a product is to wash off. On SC2 group one loadings are nearly all zeros. which captures product use properties such as how the rollball performs and on application feel. SC3 contrasts group 1 (application/product use properties) and group 5 (product efficacy). SC4 contrasts group 1 with group 4 (fragrance, freshness, smooth and soft). SC5 is the least clear, however, it contrast group 4 and some aspects of group 1 and deposits with group 5 (efficacy). Trapping of underarm hair is evident on this component with the efficacy group.

2.8 Simple Components with Variable Selection

One of the key characteristics of components that are interpretable, is that they are sparse, so that interdependent variables are found on the same loading vectors, and groupings differentiate onto separate loading vectors. A two step approach is to

1. Apply a variable selection algorithm to obtain a subset that best approximates the subspace spanned by a chosen set of principal components.
2. Find simple components using this subset

Cadima and Jolliffe (2001) describe a forward backward algorithm for variable selection. The criterion used to access the variable subsets is based on a measure of how well a subset of variables approximates a given subset of principal components, by calculating how similar are the subspaces spanned by each. The *generalized coefficient of determination* (GCD) for instance can be thought of as the average of the squared canonical correlations between the two sets of variables spanning the subspaces.

2.9 An Application of Simple Components to Large Data Sets

The enumeration of simple components may become impractical or intractable for very large data sets. For example for a problem with 500 variables the number of evaluations

Description	Group	SC1	SC2	SC3	SC4	SC5
rollball glided over skin	1	1	0	-1	0	1
ball rolled freely in pack		1	0	-1	0	1
ball did not dry out		1	0	-1	-1	1
pack did not become messy		1	0	-1	-1	0
product did not leak out		1	0	-1	-1	0
easy to apply the right amount		1	-1	-1	-1	0
easy of application		1	0	-1	0	1
ease of applying right amount		1	0	0	-1	-1
how smooth whilst applying		1	0	-1	0	1
how cold whilst applying		1	0	0	-1	-1
notice visible deposits - skin		1	0	0	-1	-1
notice deposits on clothes		1	0	0	-1	-1
overall opinion packaging		1	0	-1	0	0
dried quickly	2	1	-1	1	1	-1
how greasy whilst applying		1	-1	0	0	1
how wet whilst applying		1	-1	1	0	-1
speed of drying		1	-1	1	1	0
how sticky whilst wearing		1	-1	0	1	1
how greasy whilst wearing		1	0	0	0	1
how easy to wash off skin		0	-1	0	0	1
any irritation		0	0	0	0	0
any trapping of underarm hair		0	0	0	0	1
how often applied rollon		0	0	0	0	0
how product dosed from pack	3	0	1	-1	1	1
felt wet during application		-1	1	-1	0	1
felt sticky whilst drying		-1	1	0	-1	0
left visible deposits		-1	0	0	1	0
cold on application		-1	0	0	1	1
marked clothes		-1	0	0	1	1
waited longer than usual- drying		-1	1	-1	0	0
felt greasy		-1	1	0	0	-1
felt fresh whilst applying	4	1	0	-1	1	-1
felt smooth whilst applying		1	0	-1	1	0
left underarm soft and smooth		1	0	0	1	0
had a pleasant fragrance		1	1	-1	1	-1
fragrance lasted long enough for me		1	1	0	1	-1
how sticky whilst applying		1	-1	0	1	0
how sticky immediately after application		1	-1	0	1	0
overall opinion fragrance		1	1	-1	1	-1
strength fragrance-immediately		0	1	0	1	-1
strength fragrance- end of day		1	1	0	1	-1
gave me day-long protection - BO	5	1	1	1	0	1
gave me day-long protection- wetness		1	1	1	0	1
kept me fresh all day		1	1	1	0	1
overall opinion - effective		1	0	1	0	1
notice any perspiration		1	1	1	-1	0
overall how effective keeping you dry		1	1	1	0	1
notice any odour		1	1	1	0	0
how effective keeping free from odour		1	1	1	0	1
Var Explained		**13.30**	**3.98**	**2.91**	**2.12**	**1.90**
Var Explained (PCA)		**14.01**	**4.15**	**3.15**	**2.58**	**1.88**

Table 2.12: Simple component unscaled loadings for the sensory panel data

to calculate the first 30 components is

$$30 \times 3^k C_k^{500}.$$

The limiting factor is the huge number of combinations to consider for each component. PCA is efficient and solutions are tractable, but interpretation is exasperated by the length of the principal components, which scale the loadings so that all are extremely small. Therefore it becomes acutely difficult to to make subjective decisions as to the importance of each. When faced with data sets of this kind of magnitude the usual approaches taken are to reduce the dimensionality in stages. For instance one might use variable selection techniques to reduce the number of variables. This involves a search strategy and an appropriate evaluation function. Another approach is to apply a dimensionality reduction, such as PCA, to extract features and then work with these new features. An heuristic approach is to use a form of hierarchical clustering (of which there are many) to group variables. Then these groups are treated as near independent latent variables. In all cases the intention is that the loss of information is small and unimportant, so that the key signals in the data are preserved.

According to Thurstone's original criteria variables which are interdependent should differentiate onto different components. If the idea that interdependent variables should differentiate onto block components is used to obtain a first stage simplification of the relationships between the random variables then this structure can be used to then find simple components within each block as follows.

1. Cluster Variables into interdependent blocks using *agglomerative clustering*
2. Find principal components within each cluster
3. Take the component explaining the highest variation from within each cluster (or first q or the number explaining a minimum percentage in each cluster)
4. Create a covariance or correlation matrix from the cluster component scores
5. Extract simple components from this cluster covariance matrix.

There is then two levels of interpretation,

1. In terms of the clusters
2. In terms of the original random variables .

After selecting principal components from each cluster the covariance matrix is formed from the principal component scores. Let C_i denote the ith cluster and $\mathbf{y}_{ij}$ its jth principal component. In the general case where varying numbers of principal components

are selected within clusters the resulting covariance matrix will have a structure similar to that in Figure 2.17.

		C1	C2		C3		
	*	0	*	*	*	*	$\mathbf{y}_{11}$
C1	0	*	*	*	*	*	$\mathbf{y}_{12}$
C2	*	*	*	*	*	*	$\mathbf{y}_{21}$
	*	*	*	*	0	0	$\mathbf{y}_{31}$
C3	*	*	*	0	*	0	$\mathbf{y}_{32}$
	*	*	*	0	0	*	$\mathbf{y}_{33}$

Figure 2.17: The covariance matrix obtained from the cluster principal components. An asterisk represents a covariance. Here $\mathbf{y}_{11}, \mathbf{y}_{12}$ represent the score vectors which belong to cluster C_1. As within a cluster principal components are uncorrelated and overall the covariance matrix is sparse.

The use of clustering to group variables into relatively independent blocks is exploited by Rousson and Gasser (2004), see Section 2.10. However, in this case, information is discarded by ignoring small correlations. When applied to very large sets of multivariate data the choice of clustering algorithm will effectively determine a high proportion of the information loss. The approach in this section will allow the quantification of information loss by the choice of the number of principal components to keep for each cluster.

The following example is taken from a consumer preference study and represents a medium sized problem with seventy two variables, but illustrates the approach. Respondents were asked to rate shampoo/conditioner systems on a scale from 1-5. The study used a balanced incomplete block design, with 264 respondents, each of which rated five systems from a total of eleven, which gave a total of 1,320 observations. Each respondent scores the shampoo and conditioner separately and together as a set. There are twenty, twenty five and twenty seven questions for the conditioner, shampoo and set respectively. Five clusters were obtained by hierarchical clustering with complete linkage, using a proximity matrix based on correlation. The clusters are shown in Table 2.13. Table 2.14 shows the amount of variation explained by the first, second and third principal components from each cluster. The first principal component from the clusters were taken to represent the majority of the observed variation. A covariance matrix is constructed from these principal component scores. For the ith cluster denote this as y_{i1}, so that the covariance matrix is then

$$Sc = [\mathbf{y}_{11},\ \mathbf{y}_{21},\ \mathbf{y}_{31},\ \mathbf{y}_{41},\ \mathbf{y}_{51}]' * [\mathbf{y}_{11},\ \mathbf{y}_{21},\ \mathbf{y}_{31},\ \mathbf{y}_{41},\ \mathbf{y}_{51}]/\,(n-1)\,,$$

where n is the number of observations. This is a representation of the information in the data at the cluster level. In this example with only five clusters it is then easy to extract the simple components for this covariance matrix. In fact full enumeration is

Cluster 1	**Cluster 2**
SET___is_mild_to_hair_and_scalp	SET___does_not_build_up_on_my_hair
SET___makes_hair_cleaner_for_longer	S___is_easy_to_dispense
SET___does_not_cause_dandruff	S___has_the_right_thickness
SET___leaves_my_hair_clean	S___is_creamy
SET___does_not_irritate_scalp	S___has_the_right_whiteness
SET___does_not_make_my_hair_oily	S___is_soft_to_touch
SET___does_not_leave_my_hair_limp	S___easy_to_spread_on_hair
S___wet_hair_feels_clean_after_rinse	S___easy_to_generate_lather
S___wet_hair_is_not_sticky_after_rinse	S___has_the_right_amount_of_lather
S___scalp_is_not_irritated_after_rinse	S___lather_is_soft
S___gentle_to_hair_and_scalp	S___lather_is_creamy
C___is_easy_to_rinse_off	S___lather_is_light
C___scalp_is_not_irritated_after_rinse	S___is_easy_to_rinse_off
C___gentle_to_hair_and_scalp	
	Cluster 4
Cluster 3	SET___makes_my_hair_feel_soft
C___looks_appealing	SET___allows_my_hair_to_move_naturally
C___feels_smooth_soft_during_rinsing	SET___moisturises_my_hair
C___feels_moisturised_when_rinsing	SET___Style
C___easy_to_finger_comb_during_rinse	SET___makes_my_hair_shiny
C___easy_to_finger_comb_after_rinse	SET___makes_my_hair_more_manageable
C___wet_hair_smooth_slip_after_rinse	SET___makes_my_hair_smooth
C___wet_hair_feels_soft_after_rinse	SET___makes_hair_bouncy
C___wet_hair_does_not_feel_squeaky	SET___hair_easy_to_comb_when_wet
C___hair_is_coated_after_rinsing	SET___hair_easy_to_comb_when_dry
C___hair_is_moisturised_after_rinse	SET___nourishes_my_hair
C___hair_feels_coated	SET___does_not_make_my_hair_dry
	SET___prevents_hair_damage
	SET___leaves_coated_feel_on_dry_hair
Cluster 5	SET___leaves_my_hair_easy_to_style
C___easy_to_dispense	SET___long_lasting_style
C___has_the_right_thickness	SET___makes_hair_beautiful
C___is_creamy	SET___can_shampoo_comfortably
C___is_smooth	SET___gives_hair_a_light_finish
C___spreads_well_on_hair	S___looks_appealing
C___penetrates_well_during_application	S___feels_smooth_or_soft_during_
	S___easy_to_finger_comb_after_ri
	S___wet_hair_not_squeaky_after_r
	S___wet_hair_feels_slippery_afte
	S___wet_hair_feels_soft_hair_rin
	S___hair_is_coated_after_rinsing
	S___Coated_feel_on_wet_hair
	S___Slippery_feel_on_wet_hair

Table 2.13: The cluster structure obtained for the shampoo/conditioner consumer test. The prefix C is conditioner, S is shampoo and SET, both conditioner and shampoo.

Table 2.14: The percentage variance explained by the principal components in each cluster

	Cluster				
	1	2	3	4	5
% Var PC1	56.01	56.36	74.65	58.22	71.44
% Var PC2	7.67	9.05	5.92	9.00	10.22
% Var PC3	5.66	8.34	4.28	3.62	7.51

Table 2.15: The unscaled loadings for the five clusters

Cluster id	LV1	LV2	LV3	LV4	LV5
1	1	1	1	1	1
2	-1	-1	0	0	-1
3	1	1	1	1	0
4	1	1	1	1	1
5	1	0	0	1	0

easily tractable. Table 2.15 details the unscaled simple loadings. The clusters could be labelled as in Table 2.16. Firstly, the clustering algorithm did not differentiate cluster

Table 2.16: Possible cluster labels

Cluster id	Description
1	healthy scalp (gentle, mild, no dandruff or irritation)
2	shampoo in-use properties
3	conditioning (moisturise, soft)
4	style, damage, hair feel (manageability, dry, coated feel)
5	conditioner in-use properties

4 well, and this is reflected on the unscaled loadings. Cluster 4 is present on all components, so this is ignored in the following short discussion. The first, LV1, contrasts the shampoo in-use properties with the other clusters. LV2 contrasts health and conditioning (moisturise and soft) with the shampoo in-use properties. LV3 is an average of health and conditioning. LV4 averages health, conditioning and conditioner in-use and could be explaining the perception of a conditioner's relative health benefits. LV5 contrast health with shampoo in-use and may suggest that scalp health is influenced by the usability of the shampoo.

2.10 Related Work

The use of a priori structure present in the correlation matrix has been explored by Rousson and Gasser (2004), see Section 2.2.3. Rousson and Gasser recommend median linkage clustering. Vigneau and Qannari (2003) investigate approaches to clus-

tering around latent variables. The quantity

$$T = n \sum_{k=1}^{K} \sum_{j=1}^{p} \boldsymbol{\kappa}_{kj} \, \mathrm{cov}^2 \left(\mathbf{x}_j, \mathbf{a}_k \right)$$

is maximized where $\boldsymbol{\kappa}_{kj} = 1$ if the jth variable $\mathbf{x}_j$ belongs to the kth cluster G_k and zero otherwise. There are K clusters of variables to consider. The vector $\mathbf{a}_k$ is the latent variable associated with the kth cluster and is of length n, where n is the number of observations, and $\mathbf{a}_k$ is standardized to unit length so that $\mathbf{a}_k' \mathbf{a}_k = 1$. T can also be written as

$$T = \frac{1}{n} \sum_{k=1}^{K} \mathbf{a}_k' \mathbf{X}_k \mathbf{X}_k' \mathbf{a}_k$$

where the matrix $\mathbf{X}_k$ is formed from the variables belonging to cluster k. A solution is obtained using an iterative partitioning algorithm in which the variables are allowed to move in and out of the clusters to increase the value of T. There are three stages to the algorithm

1. **Step 1** Start with K groups of variables by random allocation or from a hierarchical clustering
2. **Step 2** In cluster G_k, the latent variable $\mathbf{a}_k$ is defined as the first standardized eigenvector of $\mathbf{X}_k \mathbf{X}_k'$
3. **Step 3** New clusters of variables are formed by assigning a variable to a group if its squared coefficient of covariance with the component of this cluster is higher than with any other eigenvector of the other clusters.

Vigneau and Qannari propose an alternative objective function when the primary goal is to capture disagreement, for instance when consumers rate their acceptability of n products. Then the quantity to maximize is

$$S = \sqrt{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \boldsymbol{\kappa}_{kj} \, \mathrm{cov} \left(\mathbf{x}_j, \mathbf{a}_k \right),$$

subject to, $\mathbf{a}_k' \mathbf{a}_k = 1$. The algorithm proceeds as previously except with the following adaptations.

1. **Step 2** In cluster G_k, it's latent variable $\mathbf{a}_k$ is set to

$$\mathbf{a}_k = \frac{\bar{\mathbf{x}}_k}{\sqrt{\bar{\mathbf{x}}_k' \bar{\mathbf{x}}_k}}$$

where $\bar{\mathbf{x}}_k$ is the centroid of cluster G_k.

2. **Step 3** New clusters are formed by moving each variable to a new group if its covariance with the standardized centroid of a group is higher than with any other standardized centroid.

As with the first partitioning approach hierarchical clustering is used which finds a decrease in S.

A recent paper by Gragn and Trendafilov (2010) describes a method to find sparse principal components using hierarchical clustering. They use hierarchical clustering but also propose an improved method to cluster the variables, called *weighted variance clustering*. They also point out that linkage methods using correlation as a measure of similarity/dissimilarity suffers from documented draw backs.

2.11 Future Work

An oblique rotation of the principal components (section 1.3.7) from the full data set can align with clusters, this is also a possibility prior to simple components. In this case the oblique factors would be used to form a correlation matrix, on which simple components can be extracted.

Alternatively, under a Gaussian assumption for the data distribution the conditional independence structure can be modelled, in the spirit of Gaussian graphical modelling (Whittaker, 1990, Edwards, 2000), which simplify the probability distribution into a clique structure. The clique structure is obtained using iterative proportional fitting via a forward or backward selection process. The difficulty arises from converting the clique structure into block components as now cliques are only conditionally independent. A simple solution may be to split cliques by distributing any shared variance between them based on the number of cliques that a variable belongs to. Perhaps a more principled approach is to apply a network clustering algorithm to extract latent variables from the cliques.

Simple components may be applied to other multivariate techniques, for instance canonical correlation analysis and partial least squares. The difficulty being that two sets of components are extracted and related simultaneously.

Chapter 3

Correlated Components

3.1 Introduction

If a set of random variables are projected onto the principal axes obtained from their covariance matrix then this will give uncorrelated scores. These also explain the maximum amount of variance. However, the interpretation of these principal axes is hampered, in general, by the large number of small non-informative weights. Specialized rotations are designed to manipulate the weights to make the axes easier to interpret. This is at the expense of redistributing variance and inducing correlation between the resulting scores. In the case of an oblique transformation the axes become dependent. According to Basilevsky (1994), when the new axes are free to take any position in the factor space, the degree of correlation allowed among factors is, in general, small because two highly correlated factors are better interpreted as only one factor. However, other orientations may be useful. For instance, if an orthogonal rotation can differentiate factors into groups, and within each group the factors are highly correlated, then within group factors can explain different aspects of that group. For example, intelligence consists of a number of faculties. There may be a group of components describing numerical skills, which are distinct from verbal skills. However, within the numerical skills group, components such as algebra, arithmetic, logic will be highly correlated with each other. An oblique rotation relaxes the condition that factors must be orthogonal and can find directions that align naturally with clusters of variables. Although orthogonal and near orthogonal axes are possible, these are not guaranteed, and the loss of orthogonality will make interpretation of each axis more difficult.

A principal component factor analysis maintains zero correlation between components by the use of standardization. If the components are not standardized then a rotation will induce correlation, (Section 1.3.8). In this chapter, orthogonal rotations, $\mathbf{z} = \mathbf{A}'\mathbf{y}$, where $\mathbf{A}'\mathbf{A} = \mathbf{I}$, of the principal components are investigated that induce maximum

covariance in the scores. These can be found subject to additional constraints on the covariance matrix. Specifically, the following $\mathbf{z}$'s are found; those that maximize

$$\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\text{cov}\left(\mathbf{z}_i,\mathbf{z}_j\right) \quad i,j=1\ldots p, \quad i\neq j \tag{3.1}$$

and those that maximize

$$\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\text{cov}\left(\mathbf{z}_i,\mathbf{z}_j\right)^2 \quad i,j=1\ldots p-1, \quad i\neq j. \tag{3.2}$$

However, it is also desirable that within groups of components that are correlated, their component loadings differentiate to explain different aspects of the group. In future, adaptations to differentiate the found axes will be considered.

Although these criteria are easy to specify, the required orthogonal rotations are not easy to find. One method might be to employ Euler angles (see, Slabaugh). For instance, if $\mathbf{E}_{ij}\left(\theta_{ij}\right)$ is the Euler matrix that rotates the ith and jth variables through an angle θ_{ij}, then the rotation matrix $\mathbf{A}$ is given by the product of these rotation matrices over all possible pairs.

$$\mathbf{A}=\prod_{i,j<i}^{p}\mathbf{E}_{ij}\left(\theta_{ij}\right). \tag{3.3}$$

In order to compare the resulting covariance matrices for a given criterion, through small increments in θ_{ij} and over all possible pairs of rotations is intractable for anything other than small problems.

Finding correlated, but orthogonal sets of components describing different aspects of a trait is one application. Another application, is to minimize the number of cross overs in a parallel coordinate plot for an arbitrary configuration of points, such as the representation obtained after using techniques such as multidimensional scaling or correspondence analysis.

3.2 Approach

Let $\mathbf{x}=(x_1,\ldots,x_p)'$ be a set of p correlated random variables with covariance matrix $\mathbf{\Sigma}_X$. If $\mathbf{x}$ is projected onto the principal axes $\mathbf{U}$, then $\mathbf{y}=\mathbf{U}'\mathbf{x}$, in which case $\mathbf{U}'\mathbf{\Sigma}_X\mathbf{U}=\mathbf{\Delta}$, where $\mathbf{\Delta}$ is a diagonal matrix of eigenvalues. As $\mathbf{U}$ is invertible, $\mathbf{U}'\mathbf{U}=\mathbf{I}$, then $\mathbf{x}=\mathbf{U}\mathbf{y}$. So without loss of generality, an orthogonal rotation can be considered from the principal axes frame of reference, $\mathbf{z}=\mathbf{A}'\mathbf{y}$, where $\mathbf{z}=(z_1,\ldots,z_p)'$. This has the corollary that any solution for $\mathbf{A}$ and $\mathbf{\Sigma}_Z=\mathbf{A}\mathbf{\Sigma}_Y\mathbf{A}'=\mathbf{A}\mathbf{\Delta}\mathbf{A}'$ can be expressed in

terms of the eigenvalues of $\mathbf{\Sigma}_X$. The eigenvalues are preserved after an orthogonal rotation giving p constraints on $\mathbf{A}$. Specific optimization criterion will impact the rotation $\mathbf{A}$ in different ways, for instance if the sum of the covariance or sum of the squared covariance is maximized. So, the problem of finding correlated orthogonal components consists of three distinct parts

1. Satisfying the eigenvalue constraints, which are universal to all optimization criteria. i.e. the eigenvalues of $\mathbf{\Sigma}_Z$ must be the same as $\mathbf{\Sigma}_X$.
2. Finding the specific sets of equations between the covariance parameters of $\mathbf{\Sigma}_Z$ that satisfy the given optimization criterion.
3. Find the orthogonal rotation $\mathbf{A}$ from $\mathbf{\Sigma}_Z$.

In general the method to find $\mathbf{A}$ is to first find the covariance matrix $\mathbf{\Sigma}_Z$ that satisfies a specific optimization criterion subject to the constraints on the eigenvalues. So, if the optimization step yields a valid solution the eigenvalues of $\mathbf{\Sigma}_Z$ are identical to those of $\mathbf{\Sigma}_X$. In which case $\mathbf{A}$ is then the matrix whose columns are the eigenvectors of $\mathbf{\Sigma}_Z$. However, $\mathbf{A}$ is the orthogonal rotation that rotates the principal components to $\mathbf{z}$. The matrix that rotates the original variables $\mathbf{x}$ to $\mathbf{z}$ is then given by $\mathbf{UA}$, as

$$\mathbf{z} = \mathbf{A}'\mathbf{y} = \mathbf{A}'\mathbf{U}'\mathbf{x} = (\mathbf{UA})'\,\mathbf{x}.$$

In the following sections the components of the problem are explored. Then specific criterion are examined individually.

Eigenvalue Constraints

An orthogonal rotation $\mathbf{z} = \mathbf{A}'\mathbf{y}$ preserves the eigenvalues of its covariance matrix, so that, $\mathbf{\Sigma}_Z$ and $\mathbf{\Sigma}_Y$ have the same eigenvalues, $\mathbf{\Delta}$.

Then the characteristic polynomials of $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_Z$ are identical.

$$\det\left(\mathbf{\Sigma}_Y - \lambda\mathbf{I}\right) = \det\left(\mathbf{\Sigma}_Z - \lambda\mathbf{I}\right).$$

This identity enables algebraic expressions to be obtained for the variance and covariance parameters of $\mathbf{\Sigma}_Z$ in terms of the eigenvalues of $\mathbf{\Sigma}_Y$. In fact using the diagonalized form of $\mathbf{\Sigma}_Y$ gives the following identity relating the covariances to the eigenvalues of $\mathbf{\Sigma}_Z$.

$$\begin{vmatrix} \lambda_1 - \lambda & 0 & \dots \\ 0 & \lambda_2 - \lambda & \dots \\ \vdots & \vdots & \ddots \end{vmatrix} \equiv \begin{vmatrix} V_{11} - \lambda & V_{12} & V_{13} & \dots \\ V_{21} & V_{22} - \lambda & V_{23} & \dots \\ V_{31} & V_{32} & V_{33} - \lambda & \dots \\ \vdots & \vdots & \vdots & \ddots \end{vmatrix} \tag{3.4}$$

where V_{ii} and V_{ij} are the variance and covariance elements of $\mathbf{\Sigma}_Z$. In general the characteristic polynomial of any $p \times p$ square matrix $\mathbf{M}$ can be written

$$f(\mu) = \mu^p - a_1\mu^{p-1} + \ldots + (-1)^p a_p$$

where $a_1 = \text{trace}(\mathbf{M})$, $a_p = \det(\mathbf{M})$ and the coefficients $a_2 \ldots a_{p-1}$ can be expressed in terms of the powers of the trace of $\mathbf{M}$.

$$\begin{aligned} a_2 &= \frac{1}{2}\left\{\text{trace}(\mathbf{M})^2 - \text{trace}(\mathbf{M}^2)\right\} \\ a_3 &= \frac{1}{6}\left\{\text{trace}(\mathbf{M})^3 - 3\text{trace}(\mathbf{M}^2)\text{trace}(\mathbf{M}) + 2\text{trace}(\mathbf{M}^3)\right\} \\ &\ldots \end{aligned}$$

These algebraic expressions can be found using Newton's identities (Mead, 1992, Kalman, 2000) as follows. Let $e_k\left(x_1, \ldots, x_n\right)$, for $k > 0$, be an elementary symmetric polynomial in n variables defined as the sum of the products of all k distinct variables. Thus

$$\begin{aligned} e_0 &= 1 \\ e_1 &= x_1 + x_2 + \ldots + x_n \\ e_2 &= \sum_{i<j} x_i x_j \\ &\ldots \\ e_n &= x_1 x_2 x_3 \ldots x_n \\ e_k &= 0 \text{ for } \quad k > n. \end{aligned}$$

Let s_k denote the power sum, $s_k = \sum_{i=1}^{n} x_i^k$, then Newton's identities can be expressed in recursive form as

$$ke_k = \sum_{i=1}^{k} (-1)^{i-1} e_{k-i} s_i, \tag{3.5}$$

giving

$$\begin{aligned} e_1 &= s_1 \\ 2e_2 &= e_1 s_1 - s_2 \\ 3e_3 &= e_2 s_1 - e_1 s_2 + s_3 \\ &\ldots \end{aligned}$$

The coefficients $a_1, a_2, \ldots a_n$ of the characteristic polynomial are equivalent to $e_1, e_2, \ldots e_n$ respectively. These elementary symmetric polynomials are connected to the trace $(\mathbf{M})$ by the fact that

$$s_i = \text{trace}\left(\mathbf{M}^i\right).$$

So, for a_2

$$\begin{aligned} a_2 = e_2 &= \frac{1}{2}\left\{e_1 s_1 - s_2\right\} \\ &= \frac{1}{2}\left\{s_1^2 - s_2\right\} \\ &= \frac{1}{2}\left\{\text{trace}\left(\mathbf{M}\right)^2 - \text{trace}\left(\mathbf{M}^2\right)\right\}. \end{aligned}$$

The characteristic polynomials (3.4) are identical, so the coefficients form a set of identities which can be solved numerically. These equations are obtained by equating the coefficients expressed in terms of their Newton identities

$$\begin{aligned} s_1 &= r_1 \\ a_1 s_1 - s_2 &= d_1 r_1 - r_2 \\ a_2 s_1 - a_1 s_2 + s_3 &= d_2 s_1 - d_1 s_2 + s_3 \\ &\ldots, \end{aligned}$$

where (a_i, s_i) and (d_i, r_i), $i = 1, \ldots, p$ are the corresponding coefficients and power sums of the characteristic polynomials. As $a_i = d_i \; \forall \; i$, these imply by forward substitution the following set of identities, $s_k = r_k \; \forall \; k$, which is

$$\text{trace}(\mathbf{\Sigma}_Z^k) = \text{trace}(\mathbf{\Delta}^k). \tag{3.6}$$

Thus equation (3.6) relates the covariance parameters of $\mathbf{\Sigma}_Z$ to the eigenvalues.

3.3 Specific Optimization Criteria

3.3.1 Maximization of the Sum of the Covariance Parameters

The first case considers the set of components that maximize the sum of the covariances between all pairs of scores by finding the orthogonal rotation, $\mathbf{A}$, such that $\mathbf{z} = \mathbf{A}'\mathbf{y}$ that maximizes

$$\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\text{cov}\left(\mathbf{z}_i, \mathbf{z}_j\right) \quad i, j = 1, \ldots, p \quad i \neq j. \tag{3.7}$$

This will give orthogonal axes, however, the scores will be correlated. As already mentioned, in general, the method for finding $\mathbf{A}$ is first to find the covariance matrix $\mathbf{\Sigma}_Z$ that satisfies the specific optimization criterion subject to the eigenvalue constraints. Then from this find $\mathbf{A}$.

The loss function, with Lagrange multipliers $\{\mu_i\}, \{\mu_{ij}\}$ corresponding to the length and orthogonal constraints on $\mathbf{A}$, and rotating from the principal components $\mathbf{y}$ (where $\text{cov}\left(\mathbf{y}\right) = \mathbf{\Delta}$), is

$$L = \frac{1}{2}\sum_i\sum_j \mathbf{a}_i'\mathbf{\Delta}\mathbf{a}_j - \frac{1}{2}\sum_i \mu_i\left(\mathbf{a}_i'\mathbf{a}_i - 1\right) - \frac{1}{2}\sum_i\sum_j \mu_{ij}\left(\mathbf{a}_i'\mathbf{a}_j\right) \quad i = 1, \ldots, p, j = 1, \ldots, p$$

Differentiating L with respect to the $\mathbf{a}_i$'s and setting to zero

$$\frac{\partial L}{\partial \mathbf{a}_i} = \frac{1}{2}\sum_{j\neq i}\mathbf{\Delta a}_j - \mu_i\mathbf{a}_i - \frac{1}{2}\sum_{j\neq i}\mu_{ij}\mathbf{a}_j = \mathbf{0}. \tag{3.8}$$

Multiplication of each of these equations by $\mathbf{a}'_k, k = 1, \ldots, p$, gives,

$$\begin{aligned}
\mathbf{a}'_k\frac{\partial L}{\partial \mathbf{a}_i} &= \frac{1}{2}\sum_{j\neq i}\mathbf{a}'_k\mathbf{\Delta a}_j - \frac{1}{2}\mu_{ik} = \mathbf{0} \quad \mathbf{k}\neq\mathbf{i} \\
\mathbf{a}'_k\frac{\partial L}{\partial \mathbf{a}_i} &= \frac{1}{2}\sum_{j\neq i}\mathbf{a}'_i\mathbf{\Delta a}_j - \mu_i = \mathbf{0} \quad \mathbf{k}=\mathbf{i}.
\end{aligned}$$

If $V_{ij} = \mathbf{a}'_i\mathbf{\Delta a}_j$ and $i, j, k = 1 \ldots p$, then,

$$\sum_{j\neq i, k\neq i} V_{kj} - \mu_{ik} = 0 \qquad i, j, k = 1, \ldots, p \tag{3.9}$$

$$\frac{1}{2}\sum_{j\neq i} V_{ij} - \mu_i = 0 \qquad i = 1, \ldots, p, \tag{3.10}$$

Let $\mathbf{\Sigma}_Z$ denote the covariance matrix of $\mathbf{A'y}$. The equations generate a set of constraints on $\mathbf{\Sigma}_Z$ which ensure that its row (or column) sums are identical. To see this, $\mu_{ij} = \mu_{ji}$, then from (3.9), and as $V_{ik} = V_{ki}$,

$$\begin{aligned}
\sum_{j\neq i, k\neq i} V_{kj} &= \sum_{j\neq k, i\neq k} V_{ij} \\
\sum_{j\neq i, k\neq i} V_{kj} + V_{ki} &= \sum_{j\neq k, i\neq k} V_{ij} + V_{ik}.
\end{aligned} \tag{3.11}$$

The left hand side of equation (3.11) is the ith row sum, and the right hand side is the kth row sum. Hence, all row sums are identical.

Denoting the row sum as t the sum of all the elements can be expressed in terms of the p row sums. Then the sum of the covariances at the maximum is

$$M = \sum_{j>i} V_{ij} = \frac{pt - \text{trace}(\mathbf{\Sigma}_Z)}{2} = \frac{pt - \text{trace}(\mathbf{\Delta})}{2}. \tag{3.12}$$

Further, from (3.9) and (3.10), $\mu_i = t - V_{ii}$ and $\mu_{ij} = t - V_{ij}$. Substituting back into (3.8) for μ_i and μ_{ij}.

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{a}_i} &= \sum_{j\neq i}\mathbf{\Delta a}_j - (t - V_{ii})\,\mathbf{a}_i - \sum_{j\neq i}(t - V_{ij})\,\mathbf{a}_j = \mathbf{0} \\
&= \sum_{j\neq i}\mathbf{\Delta a}_j - t\mathbf{a}_i + V_{ii}\mathbf{a}_i - \sum_{j\neq i} t\mathbf{a}_j + \sum_{j\neq i} V_{ij}\mathbf{a}_j = \mathbf{0} \\
&= \sum_{j\neq i}\mathbf{\Delta a}_j + \left(\sum_{j\neq i} V_{ij}\mathbf{a}_j + V_{ii}\mathbf{a}_i\right) - t\mathbf{a}_i - \sum_{j\neq i} t\mathbf{a}_j = \mathbf{0}.
\end{aligned}$$

Now

$$\sum_{j \neq i} V_{ij}\mathbf{a}_j + V_{ii}\mathbf{a}_i = \sum_{j=1}^{p} V_{ij}\mathbf{a}_j,$$

so,

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{a}_i} &= \sum_{j \neq i} \mathbf{\Delta}\mathbf{a}_j + \sum_j V_{ij}\mathbf{a}_j - t\mathbf{a}_i - \sum_{j \neq i} t\mathbf{a}_j = \mathbf{0} \\
&= \sum_{j \neq i} \mathbf{\Delta}\mathbf{a}_j + \sum_j V_{ij}\mathbf{a}_j - \left(t\mathbf{a}_i + \sum_{j \neq i} t\mathbf{a}_j \right) = \mathbf{0}
\end{aligned}$$

and

$$t\mathbf{a}_i + \sum_{j \neq i} t\mathbf{a}_j = \sum_j t\mathbf{a}_j,$$

so

$$\frac{\partial L}{\partial \mathbf{a}_i} = \sum_{j \neq i} \mathbf{\Delta}\mathbf{a}_j + \sum_j V_{ij}\mathbf{a}_j - \sum_j t\mathbf{a}_j = \mathbf{0}.$$

Now, summing over all the partial derivatives,

$$\begin{aligned}
\sum_i \frac{\partial L}{\partial \mathbf{a}_i} &= \sum_i \sum_{j \neq i} \mathbf{\Delta}\mathbf{a}_j + \sum_i \sum_j V_{ij}\mathbf{a}_j - \sum_i \sum_j t\mathbf{a}_j = \mathbf{0} \\
&= (p-1)\,\mathbf{\Delta} \sum_j \mathbf{a}_j + t \sum_j \mathbf{a}_j - pt \sum_j \mathbf{a}_j = \mathbf{0},
\end{aligned}$$

as

$$\sum_i \sum_j V_{ij}\mathbf{a}_j = t \sum_j \mathbf{a}_j,$$

due to symmetry, $V_{ij} = V_{ji}$. This gives,

$$[(p-1)\,\mathbf{\Delta} - (p-1)\,t\,\mathbf{I}] \sum_j \mathbf{a}_j = \mathbf{0}.$$

This shows that the row sum t is an eigenvalue of $\mathbf{\Delta}$ (one of it's diagonal values), and $\sum_j \mathbf{a}_j$ is the corresponding eigenvector. In order to maximize the sum of the covariance each row sum must be equal to the largest value of $\mathbf{\Delta}$, λ_1,

$$\sum_j V_{ij} = \lambda_1 \quad \forall \quad i = 1 \ldots p.$$

Substituting t back into (3.12) the maximum in terms of the eigenvalues is

$$M = \frac{p\lambda_1 - \sum_{i=1}^{p} \lambda_i}{2}.$$

Identifiability of Solutions

To maximize $\sum_{j>i} \text{cov}\,(\mathbf{z}_i, \mathbf{z}_j)$ the following sets of equations must be satisfied.

$$\begin{aligned} \sum_{j=1}^{p} V_{ij} - \lambda_1 &= 0 \qquad & i = 1, \ldots, p \\ \text{trace}(\boldsymbol{\Sigma}_y^k) &= \text{trace}(\boldsymbol{\Delta}^k) \qquad & k = 1, \ldots, p. \end{aligned} \tag{3.13}$$

There are p independent row sums and p eigenvalue identities. Finding the $p+p(p-1)/2$ covariance parameters is equivalent to finding the orthogonal rotation $\mathbf{A}$. This leaves a requirement of

$$p + \frac{p(p-1)}{2} - 2p = \frac{p(p-3)}{2}$$

additional constraints in order to identify a single solution (or no solution). In the case of three variables no additional constraints are necessary. For larger p how these additional constraints are chosen is important. For instance, covariance values could be fixed or constrained in order to obtain a solution conforming to a prescribed pattern in the covariance matrix. The number to obtain a unique solution rapidly increases, so a data set with 50 variables, requires 1,175 additional constraints. Importantly, if a single solution is required from the set of solutions defined by (3.13), it may or may not be contained within the solution subset that is constrained by the chosen additional constraints.

If the rotation constraints, eigenvalue constraints and the additional constraints are combined it is possible to determine the parameters of $\boldsymbol{\Sigma}_Z$ by solving the resulting set of equations numerically. This is consistent with the number of parameters to find in $\mathbf{A}$ as there are p length constraints and $p(p-1)/2$ orthogonality constraints. An approach to finding a unique solution is to take (3.13) and additional constraints and solve the resulting set of non-linear equations using a standard method such as Newton-Raphson. However, because there is no guarantee that a single solution can be found and the method is sensitive to the choice of starting point, it becomes difficult to know if there is no solution or a bad starting point has been selected. To address this a perturbation algorithm has been developed which is not dependent on selecting starting values and will indicate if a valid solution was found or not. This is outlined in the next section. Another approach is to incorporate extra constraints explicitly by including them in the loss function. This line of research has not yet been explored extensively and is a topic for future work.

A New Perturbation Algorithm

In order to obtain a valid solution for the constraints placed on the covariance structure, the solution must have the same eigenvalue structure as the original covariance matrix.

However, the variance parameters are free to take any value provided the eigenvalue constraints are not violated. The algorithm exploits this by adjusting the variance parameters of the current solution for the covariance parameters. This avoids solving the characteristic polynomial explicitly (equation (3.13)) using a Newton-Raphson method. The following loss function is minimized,

$$\sum_{i=1}^{p} \{\lambda_i - \eta_i\}^2 \tag{3.14}$$

where λ_i is the ith largest eigenvalue of $\mathbf{\Sigma}_X$ and η_i the ith largest eigenvalue of the current solution. The eigenvalues represent the maximum and minimum values the variances can take. Hence the value of a V_{ii} must lie within the range $[\lambda_p\ \lambda_1]$ and can be ordered $V_{pp} \leq \ldots \leq V_{11}$. The algorithm proceeds by taking each V_{ii} in turn and perturbing it by a small increment and also decrement. Then the row sum equations are solved to yield updated covariance estimates. The eigenvalues of the updated covariance matrix are used to calculate the loss, equation (3.14). At each iteration all V_{ii} are taken in turn and the one which yields the smallest loss is taken forward. This process is repeated until the algorithm converges to the original eigenvalue structure. The algorithm is summarized in **Algorithm 2**. In practice the variances are perturbed

Input: Eigenvalues, $\lambda_1, \ldots, \lambda_p$ of $\mathbf{\Sigma}_X$
Output: V
Initialize by setting $V_{ii} = \lambda_i$ and $V_{ij} = 0\ \forall\ i, j$
forall the δ *in [0.1 0.01 0.001 0.0001 0.00001 ...]* **do**
  repeat
    forall the V_{ii} **do**
      Perturb current V_{ii} by $\pm\delta$ to give 2^p solution sets
      Calculate the eigenvalues for each solution set $\{\eta\}$
      Select the solution set that minimizes the loss $\sum_i \{\lambda_i - \eta_i\}^2$
      Update all V_{ii} by solving $\sum_i V_{ij} = \lambda_1$ for $i, j := 1 \ldots p$
    end
  until *no further improvement*;
end

Algorithm 2: Iterative algorithm to find the rotation which maximizes the sum of the covariance elements

with increasing refinement.

3.3.2 Maximization of the Squared Sum of the Covariance Parameters

Changing the optimisation criterion will generate different kinds of structures. In the first instance

$$\max \sum \text{cov}\,(\mathbf{z}_i, \mathbf{z}_j)$$

will penalize negative covariance. If the criterion is modified slightly to maximise the sum of the squared covariance between the components

$$\max \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \operatorname{cov}\left(\mathbf{z}_i, \mathbf{z}_j\right)^2 \quad i, j = 1, \ldots, p \quad i \neq j \tag{3.15}$$

then absolute covariance elements can be large. The modified loss function is

$$L = \frac{1}{2} \sum_{i} \sum_{j, j \neq i} \left(\mathbf{a}_i' \boldsymbol{\Delta} \mathbf{a}_j\right)^2 - \frac{1}{2} \sum_{i} \mu_i \left(\mathbf{a}_i' \mathbf{a}_i - 1\right) - \sum_{i} \sum_{j, j \neq i} \mu_{ij} \left(\mathbf{a}_i' \mathbf{a}_j\right) \quad i, j = 1, \ldots, p.$$

Differentiating L with respect to the $\mathbf{a}_i$'s and setting to zero gives

$$\frac{\delta L}{\delta \mathbf{a}_i} = \sum_{j \neq i} \left(\mathbf{a}_i' \boldsymbol{\Delta} \mathbf{a}_j\right) \boldsymbol{\Delta} \mathbf{a}_j - \mu_i \mathbf{a}_i - \sum_{j \neq i} \mu_{ij} \mathbf{a}_j = \mathbf{0}.$$

Multiplication of each of these equations by $\mathbf{a}_k'$, $k = 1, \ldots, p$, generates the set of equations

$$\begin{aligned} \sum_{j \neq i} V_{ij} V_{kj} - \mu_{ik} &= 0 \quad k \neq i \\ \sum_{j \neq k} V_{kj}^2 - \mu_k &= 0 \quad k = i. \end{aligned}$$

Equating the equations sharing a Lagrange multiplier, i.e. μ_{ki} and μ_{ik} in the above, reveals that

$$V_{kk} V_{ki} = V_{ii} V_{ik},$$

for all i and k, and as $V_{ik} = V_{ki}$ by symmetry in the covariance matrix, then

$$V_{ik} \left(V_{kk} - V_{ii}\right) = 0.$$

This implies that if $abs(V_{ik}) > 0$ then the variance parameters V_{ii} and V_{kk} must be identical. In essence, the constraints imposed on the variance parameters are a consequence of the orthogonality constraints placed on the rotation matrix $\mathbf{A}$, which relate to the μ_{ij}'s from the Lagrange multipliers. Also,

$$\mu_i = \sum_{j=1}^{p} V_{ij}^2 \quad \forall \quad i = 1, \ldots, p.$$

However, this does not help solve for the parameters as the μ_i's are unknown. There are $p\left(p-1\right)/2$ covariance parameters to find. At the maximum, the variance parameters are equal if the corresponding covariance parameters are not zero. For the general case where none are zero the variance parameter is

$$V = \frac{1}{p} \operatorname{trace}\left(\boldsymbol{\Sigma}_X\right) = \frac{1}{p} \sum_{i=1}^{p} \lambda_i.$$

In this case the characteristic equation relates the eigenvalues, $\mathbf{\Delta}$, to the covariance parameters of $\mathbf{\Sigma}_Z$. For the three variable case

$$\begin{vmatrix} V-\lambda & V_{12} & V_{13} \\ V_{12} & V-\lambda & V_{23} \\ V_{13} & V_{23} & V-\lambda \end{vmatrix} = \begin{vmatrix} \lambda-\lambda_1 & 0 & 0 \\ 0 & \lambda-\lambda_2 & 0 \\ 0 & 0 & \lambda-\lambda_3 \end{vmatrix}. \tag{3.16}$$

Solutions can be found numerically by solving the sets of identities generated from these characteristic polynomials, or using the identities given in equation (3.6).

To summarize the problem of finding correlated components where the sum of their squared covariances is maximized,

Optimization constraint

$$V_{ij}\left(V_{ii}-V_{jj}\right)=0 \quad \forall \quad i,j=1\ldots p$$

When $V_{ij} \neq 0$ all the variance parameters are identical and then,

$$V_{ii}=\frac{1}{p}\text{trace}(\mathbf{\Delta}) \quad \forall \quad i=1\ldots p.$$

If V_{ij} is zero then V_{ii} and V_{jj} are not necessarily identical and may be extra parameters that need to be estimated. This can only happen if all of the covariance parameters involving i or j are also zero, that is, when variables are in independent groups.

Eigenvalue Constraints

$$\text{trace}\left(\mathbf{\Sigma}_Z^k\right) \equiv \text{trace}\left(\mathbf{\Delta}^k\right) \; \forall \; k=1\ldots p$$

Additional Constraints

There are $1+p(p-3)/2$ extra non-dominated constraints required to identify a unique solution. That is, there are $p+p(p-1)/2$ covariance parameters to identify. However, there are p rotation constraints and $p-1$ optimization constraints, in that only a single variance parameter is required for the general case. As pointed out in Section 3.3.1, where these are placed, their magnitude and form, will impact on whether a unique solution can be found within the set of valid solutions defined by the optimization and rotation constraints.

Finding Solutions

Unlike the first optimization criterion, equation (3.7), where the variance parameters are free to take any value, the variance parameters have to equal $\sum_{i=1}^{p}\lambda_i/p$. Although a general Euler angle approach is intractable, it is used here to rotate pairs of axes until

their variances are equal. For example, to rotate the first and fourth axes, a general rotation of the form

$$\begin{bmatrix} \cos\theta & 0 & 0 & -\sin\theta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \sin\theta & 0 & 0 & \cos\theta \end{bmatrix} \tag{3.17}$$

is applied. There are

$$C_2^p = \frac{p(p-1)}{2}$$

pairwise rotations to consider. The algorithm is outlined in pseudo code in **Algorithm 3**. In practice, the Euler angle approach is used to find a good starting solution to feed into a non-linear equation solver. In which case, the characteristic equation identities (3.16) or the trace identities are used (3.6), combined with the variance constraint and any additional constraints.

Input: Sample covariance matrix, $\mathbf{S}_0 = [V_{ij}]$, number of iterations N
Output: Rotated Covariance matrix, $\mathbf{S}^*$
$L_{\min} :=$ var (trace ($\mathbf{S}_0$))
for *cnt := 1 to N* **do**
  Select a pair of variables, i, j at random
  if $V_{ij} \neq 0$ **then**
    $\theta := \frac{1}{2} \arctan\left(\frac{V_{ii}-V_{jj}}{2V_{ij}}\right)$
  else
    $\theta := \pi/2$
  end
  Construct $\mathbf{R}$ for θ as in equation (3.17)
  $\mathbf{S}_{\text{cnt}} := \mathbf{R}'\mathbf{S}_{(\text{cnt}-1)}\mathbf{R}$
  $L_{\text{cnt}} :=$ var (trace ($\mathbf{S}_{\text{cnt}}$))
  if $L_{\text{cnt}} < L_{\min}$ **then**
    $\mathbf{S}^* := \mathbf{S}_{\text{cnt}}$
    $L_{\min} := L_{\text{cnt}}$
  end
end

Algorithm 3: Euler rotation to find a solution where the variance parameters are identical. Orthogonality of the axes and the eigenvalues are preserved

3.4 Correlated Component Analysis of the Deodorant Data

The deodorant data introduced in Section 1.1.3, which consists of 49 variables was re-analysed. Five principal components were extracted and rotated using both the sum of the covariance, equation (3.7) and the sum of the squared covariance, equation (3.15) criterion. The resulting correlation between the scores are shown in Table 3.1.

a.

Z1	1				
Z2	0.28	1			
Z3	0.46	0.12	1		
Z4	0.33	0.01	0.19	1	
Z5	-0.59	-0.13	-0.38	-0.20	1
	Z1	Z2	Z3	Z4	Z5

b.

Z1	1				
Z2	-0.14	1			
Z3	0.67	-0.41	1		
Z4	0.30	0.00	0.38	1	
Z5	-0.48	0.06	-0.51	-0.33	1
	Z1	Z2	Z3	Z4	Z5

Table 3.1: a. The correlation between the five rotated principal components using the sum of the covariance criterion (3.7). b. Correlations using the sum of the squared covariance (3.15)

The correlation structure in table a. is more restrictive than that of b. As all the first off diagonal correlations are all positive this induces a negative correlation between Z2 and Z5. The structure of b. in this example, results in the correlations between Z2 and Z5, and Z2 and Z4 to be near zero. Table b. also has higher absolute correlations and is more differentiated. Tables 3.2 and 3.3 show the component loadings and the correlation of the variables with the component scores. Components Z2 and Z3 have a correlation of -0.41. Z2 could be summarized as describing how effective a product is at keeping the respondent dry and the lack of negative drivers, for instance, *didn't feel greasy*, *didn't feel sticky* and *didn't leave deposits*, *didn't mark clothes*, during drying and on application. Z3 could be summarized as describing the speed of drying properties of the products in the study. Taken together, Z2 and Z3 can be considered as describing different aspects of product properties whilst drying and on application, where the speed of drying (Z3) is anti-correlated with whilst drying and on application properties.

Z1 describes the sensory properties in the underarm on application e.g. *smooth, soft, fresh*. Z1 also shares some of the properties of Z2, e.g. *did not leave visible deposits*, *sticky whilst wearing* and also with Z5, e.g. *effective against wetness and odour*. Z4 describes pack properties and performance, however the large correlation between the variables with this component are negative, which confuses the interpretation as Z4 is anti-correlated with Z5. Z5 describes product efficacy, and is anti-correlated with Z1, Z3 and Z4, perhaps indicating that a perception of good sensory and skin care properties also goes hand in hand with a perception of lower efficacy and visa versa.

Variables	**CC1**	**CC2**	**CC3**	**CC4**	**CC5**
rollball glided over skin	0.22	0.04	0.09	-0.23	-0.11
felt fresh whilst applying	0.24	-0.01	0.05	-0.04	-0.01
felt smooth whilst applying	0.24	-0.07	0.04	-0.10	-0.07
didn't feel wet during application	-0.16	-0.12	0.40	0.03	0.13
didn't feel sticky whilst drying	0.12	-0.41	0.08	0.08	-0.07
dried quickly	-0.10	-0.24	0.27	0.13	0.09
left underarm soft and smooth	0.12	-0.03	0.13	0.03	0.02
had a pleasant fragrance	0.32	0.10	0.03	0.12	0.30
fragrance lasted long enough for me	0.21	0.24	0.09	0.31	0.23
did not leave visible deposits	0.09	0.04	0.12	-0.04	-0.08
gave me day-long protection - BO	0.00	0.19	0.13	0.20	-0.21
gave me day-long protection- wetness	-0.05	0.13	0.12	0.22	-0.31
kept me fresh all day	0.02	0.15	0.12	0.23	-0.24
not cold on application	0.02	0.07	0.20	-0.06	0.00
didn't mark clothes	0.04	0.00	0.12	-0.05	-0.10
didn't wait longer than usual- drying	-0.09	-0.19	0.32	0.08	0.12
didn't feel greasy	0.12	-0.17	0.11	-0.04	-0.05
ball rolled freely in pack	0.18	0.05	0.11	-0.25	-0.10
ball did not dry out	0.20	0.09	0.14	-0.33	-0.14
pack did not become messy	0.08	0.04	0.10	-0.14	-0.06
product did not leek out	0.06	0.01	0.06	-0.09	-0.05
easy to apply the right amount	0.03	0.13	0.31	-0.23	0.06
easy of application	0.15	0.05	0.10	-0.17	-0.08
how product dosed from pack	0.12	-0.01	-0.12	-0.03	-0.06
ease of applying right amount	0.00	0.14	0.30	-0.18	0.04
how smooth whilst applying	0.16	0.00	0.04	-0.10	-0.08
how sticky whilst applying	0.20	-0.38	-0.05	0.08	-0.09
how greasy whilst applying	0.13	-0.15	0.02	-0.06	-0.10
how wet whilst applying	-0.16	-0.08	0.29	0.07	0.13
how cold whilst applying	-0.01	0.01	0.17	-0.02	0.02
how sticky immediately after application	0.19	-0.42	-0.06	0.11	-0.07
speed of drying	-0.09	-0.15	0.19	0.09	0.04
speed of drying compared to usual	-0.10	-0.10	0.20	0.10	0.06
how sticky whilst wearing	0.20	-0.20	-0.08	0.08	-0.13
how greasy whilst wearing	0.13	-0.08	0.00	-0.01	-0.11
overall opinion - effective	0.15	0.08	0.15	0.20	-0.21
notice any perspiration	-0.03	0.06	0.02	0.15	-0.21
overall how effective keeping you dry	0.00	0.08	0.06	0.19	-0.29
how keeping you dry compares to usual	0.01	0.02	0.05	0.16	-0.14
notice any odour	0.02	0.05	-0.01	0.10	-0.11
how effective keeping free from odour	0.08	0.09	0.01	0.16	-0.21
how free from odour compares to usual	0.03	0.05	0.05	0.14	-0.12
overall opinion fragrance	0.45	0.11	0.01	0.24	0.41
strength fragrance-immediately	0.07	0.07	-0.01	0.09	0.10
strength fragrance- end of day	0.11	0.11	0.01	0.14	0.07
notice visible deposits - skin	0.02	-0.02	0.03	0.00	-0.06
notice deposits on clothes	0.01	-0.01	0.02	0.00	-0.05
how easy to wash off skin	0.04	-0.03	0.02	-0.02	0.01

Table 3.2: Loadings for the correlated component scores obtained by maximizing the sum of the covariance squared.

Variables	**Z1**	**Z2**	**Z3**	**Z4**	**Z5**
rollball glided over skin	0.78	0.60	0.04	-0.77	0.09
felt fresh whilst applying	0.93	0.66	0.44	-0.80	0.09
felt smooth whilst applying	0.87	0.71	0.33	-0.80	-0.03
didn't feel wet during application	0.20	0.67	0.68	-0.83	-0.16
didn't feel sticky whilst drying	0.49	0.83	0.65	-0.69	-0.34
dried quickly	0.23	0.72	0.78	-0.73	-0.25
left underarm soft and smooth	0.80	0.78	0.70	-0.88	0.10
had a pleasant fragrance	0.68	-0.02	0.51	-0.44	0.14
fragrance lasted long enough for me	0.65	0.13	0.69	-0.31	0.56
did not leave visible deposits	0.84	0.82	0.47	-0.84	0.30
gave me day-long protection - BO	0.66	0.68	0.61	-0.38	0.77
gave me day-long protection- wetness	0.61	0.75	0.60	-0.35	0.71
kept me fresh all day	0.69	0.72	0.63	-0.39	0.72
not cold on application	0.65	0.73	0.49	-0.95	0.24
didn't mark clothes	0.72	0.90	0.45	-0.85	0.22
didn't wait longer than usual- drying	0.26	0.71	0.75	-0.80	-0.22
didn't feel greasy	0.65	0.86	0.55	-0.86	-0.19
ball rolled freely in pack	0.71	0.57	-0.02	-0.78	0.06
ball did not dry out	0.68	0.53	-0.09	-0.74	0.10
pack did not become messy	0.70	0.64	0.08	-0.84	0.11
product did not leak out	0.70	0.69	0.09	-0.83	0.07
easy to apply the right amount	0.48	0.49	0.17	-0.94	0.07
easy of application	0.79	0.64	0.12	-0.81	0.16
how product dosed from pack	0.49	-0.10	-0.43	0.28	0.08
ease of applying right amount	0.50	0.56	0.26	-0.94	0.17
how smooth whilst applying	0.87	0.71	0.23	-0.77	0.11
how sticky whilst applying	0.58	0.78	0.59	-0.58	-0.36
how greasy whilst applying	0.71	0.84	0.37	-0.75	-0.21
how wet whilst applying	0.07	0.58	0.73	-0.72	-0.15
how cold whilst applying	0.52	0.78	0.63	-0.95	0.09
how sticky immediately after application	0.52	0.75	0.60	-0.54	-0.41
speed of drying	0.23	0.75	0.79	-0.71	-0.18
speed of drying compared to usual	0.20	0.70	0.82	-0.71	-0.11
how sticky whilst wearing	0.78	0.81	0.57	-0.51	-0.08
how greasy whilst wearing	0.86	0.85	0.43	-0.68	0.05
overall opinion - effective	0.82	0.81	0.69	-0.61	0.51
notice any perspiration	0.53	0.70	0.54	-0.17	0.72
overall how effective keeping you dry	0.65	0.77	0.58	-0.32	0.68
how keeping you dry compares to usual	0.66	0.80	0.74	-0.40	0.57
notice any odour	0.65	0.60	0.54	-0.14	0.76
how effective keeping free from odour	0.76	0.68	0.56	-0.30	0.71
how free from odour compares to usual	0.72	0.74	0.71	-0.40	0.63
overall opinion fragrance	0.69	0.01	0.58	-0.39	0.16
strength fragrance-immediately	0.36	-0.36	0.42	0.05	0.41
strength fragrance- end of day	0.67	0.09	0.60	-0.17	0.64
notice visible deposits - skin	0.74	0.96	0.51	-0.69	0.26
notice deposits on clothes	0.71	0.96	0.50	-0.65	0.31
how easy to wash off skin	0.65	0.58	0.36	-0.87	-0.36

Table 3.3: The correlation of the variables with the correlated component scores. The sum of the squared covariance was maximized.

3.5 An Improved Parallel Coordinate Plot for a Rotatable Configuration of Points

A popular visualization method for multivariate data is the parallel coordinate plot, which displays variables or dimensions as a set of parallel axes, so that individual observations can be compared graphically on a two dimensional plot. Figure 3.1 shows a parallel coordinate plot for a five dimensional representation, obtained from a non-metric MDS (see Section 1.2), for the deodorant data introduced in Section 1.1.3. The configuration represents the forty nine variable attributes embedded in a five dimensional space. This was obtained from a non-metric MDS using a Euclidean proximity matrix. In the case of a non-metric MDS, the axes are arbitrary and have no specific meaning. Hence, the configuration may be rotated to align with any chosen set of axes. If the configuration is rotated so that the induced correlations are maximized then the new axes can be used for a parallel coordinate plot and will minimize the number of cross overs between pairs of axes. One limitation of a parallel coordinate plot is that axes can only be viewed in a pairwise fashion, i.e. if the axis for Dim2 is drawn next to Dim1, then the axis for Dim3 cannot be visualised compared to Dim1 without re-ordering the axes. Therefore, the new plot axes are reordered so that the covariance between pairs of axes is maximized. Finally as the rotation is orthogonal the configuration will remain in an interpretable coordinate system.

A suitable target pattern for the covariance matrix is a banded structure, Table 3.4. Successive pairs of variables have a substantial covariance, but not other pairs. In this case there are $2p-1$ parameters to find, the remaining parameters being constrained to zero. As there are p eigenvalue constraints and p non redundant constraints imposed by the covariance constraints, equation (3.1) or (3.2), there must be either a unique solution or no solution for this structure. In general the eigenvalue constraint will not be met for this structure and so there will not be a solution. As the objective

Table 3.4: The target structure for a parallel coordinate plot where each successive variable has the maximum covariance with its neighbour.

$$\begin{pmatrix} v_{11} & v_{12} & & & & & \\ v_{12} & v_{22} & v_{23} & & & & \\ & v_{23} & v_{33} & v_{34} & & & \\ & & v_{34} & v_{44} & v_{45} & & \\ & & & v_{45} & v_{55} & v_{56} & \\ & & & & v_{56} & v_{66} & v_{67} \\ & & & & & v_{67} & v_{77} \end{pmatrix}$$

of equation (3.1) or (3.2) cannot be met a solution is found which maximizes these criterion without imposing any additional constraints on the covariance parameters. This ensures a solution may be found, which may not unique. This solution can then

be reordered so that the off diagonal elements of the final correlation matrix are as large as possible. This is equivalent to reordering the axes in the parallel coordinate plot. One way to do this is to reorganise to Robinson form (Brusco et al., 2007).

Starting from the principal components of the $N \times p$ configuration $\mathbf{X}$.

$$\mathbf{Y} = \mathbf{X}\,\mathbf{U},$$

where $\mathbf{U}$ are the eigenvectors of $\mathbf{S}_X = \mathbf{X}'\mathbf{X}/(N-1)$. Let $\mathbf{A}$ be the orthogonal rotation obtained from maximizing Equation (3.7) or (3.15). Then,

$$\mathbf{Z} = \mathbf{Y}\,\mathbf{A}$$

is the new configuration. $\mathbf{A}$ is the matrix whose columns are the eigenvectors of $\mathbf{S}_Z = \mathbf{Z}'\mathbf{Z}/(N-1)$.

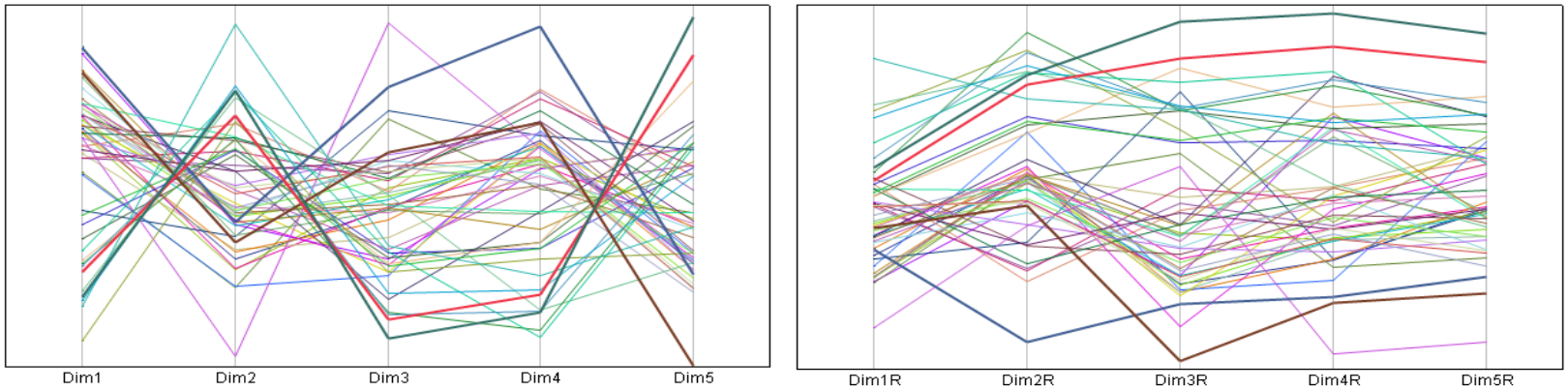
If the covariance matrix $\mathbf{S}_Z$ is permuted to Robinson form by the permutation matrix $\mathbf{P}$, where $\mathbf{PP} = \mathbf{I}$, then the eigenvectors of $\mathbf{PS}_Z\mathbf{P}$ are a permutation of the rows of $\mathbf{A}$. Let the permutation of $\mathbf{A}$ be denoted by $\mathbf{A}^* = \mathbf{AP}$, then the final configuration is given by

$$\mathbf{Z}^* = \mathbf{Y}\,\mathbf{A}^*.$$

As an example, returning to the sensory data introduced in Section 1.1.3. An MDS on the data produced a five dimensional representation space. One advantage of using a parallel coordinate plot in this way is that interrogating a dimensional representation greater than two or three will retain a high proportion of the variation, which is otherwise lost. Table 3.5 shows the correlation between the dimensions in the initial configuration and Table 3.6 the correlations between the rotated dimensions. As expected using the sum of the covariance criteria, equation (3.7), has pushed the correlations to be positive. In Figure 3.1 the two plots are compared to illustrate the utility of the new plot to identify potential outliers. While, Figure 3.2 identifies a small cluster of variables consisting of, *felt wet during application*, *felt sticky whilst drying*, *left visible deposits*, *marked clothes*, and *felt greasy*. On the rotated plot it has become easier to differentiate this cluster from the main cluster of variables.

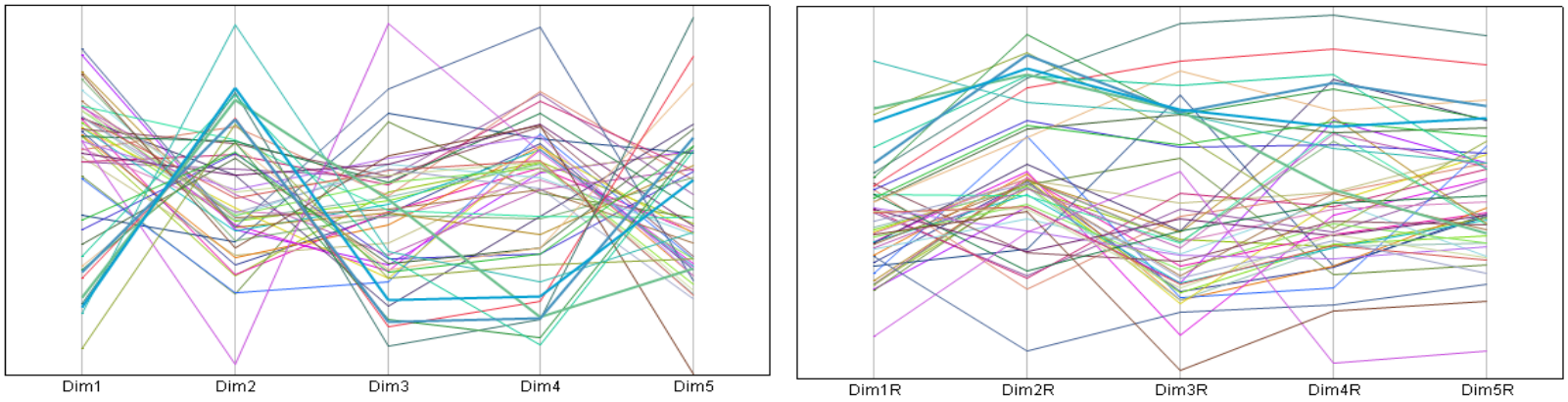
Dim1	1				
Dim2	-0.66	1			
Dim3	0.51	-0.53	1		
Dim4	0.73	-0.57	0.66	1	
Dim5	-0.48	0.42	-0.59	-0.53	1
	Dim1	Dim2	Dim3	Dim4	Dim5

Table 3.5: The correlation between the dimensions in the initial representation space.



11/0 Cols / 10/4 Rows	Attribute	Dim1	Dim2	Dim3	Dim4	Dim5	Dim1R	Dim2R	Dim3R	Dim4R	Dim5R
1	overall opinion - effective	0.55	-0.12	0.49	0.67	-0.42	-0.68	-2.56	-1.19	-1.49	-1.46
2	overall opinion fragrance	0.11	-0.56	0.74	0.16	-0.33	-2.18	-0.82	0.52	-2.26	-2.54
3	overall opinion packaging	0.43	-0.18	0.24	0.27	-0.99	-0.26	-0.52	-1.89	-1.58	-1.73
4	any irritation	-0.53	0.24	-0.41	-0.45	0.95	0.65	1.32	1.85	1.95	2.07
5	how often applied rollon	-0.66	0.32	-0.49	-0.52	1.18	0.89	1.46	2.31	2.39	2.55
6	felt wet during application	-0.63	0.30	0.07	-0.51	-0.34	2.11	1.52	1.26	0.05	-0.63
7	felt sticky whilst drying	-0.87	0.22	-0.23	-0.30	-0.29	1.99	1.84	0.97	-0.47	0.85
8	left visible deposits	-0.50	0.16	-0.39	-0.59	0.41	0.80	2.10	1.21	1.41	1.21
9	marked clothes	-0.51	0.24	-0.39	-0.52	0.45	1.08	1.79	1.23	1.48	1.42
10	felt greasy	-0.67	0.34	-0.31	-0.42	0.20	1.86	1.61	1.27	0.91	1.23

Figure 3.1: Parallel coordinate plot of the MDS configuration showing the configuration before (left) and after rotation. The highlighted variables are potential outliers. These are easier to differentiate on the updated plot.



11/0 Cols / 10/5 Rows	Attribute	Dim1	Dim2	Dim3	Dim4	Dim5	Dim1R	Dim2R	Dim3R	Dim4R	Dim5R
1	overall opinion - effective	0.55	-0.12	0.49	0.67	-0.42	-0.68	-2.56	-1.19	-1.49	-1.46
2	overall opinion fragrance	0.11	-0.56	0.74	0.16	-0.33	-2.18	-0.82	0.52	-2.26	-2.54
3	overall opinion packaging	0.43	-0.18	0.24	0.27	-0.99	-0.26	-0.52	-1.89	-1.58	-1.73
4	any irritation	-0.53	0.24	-0.41	-0.45	0.95	0.65	1.32	1.85	1.95	2.07
5	how often applied rollon	-0.66	0.32	-0.49	-0.52	1.18	0.89	1.46	2.31	2.39	2.55
6	felt wet during application	-0.63	0.30	0.07	-0.51	-0.34	2.11	1.52	1.26	0.05	-0.63
7	felt sticky whilst drying	-0.87	0.22	-0.23	-0.30	-0.29	1.99	1.84	0.97	-0.47	0.85
8	left visible deposits	-0.50	0.16	-0.39	-0.59	0.41	0.80	2.10	1.21	1.41	1.21
9	marked clothes	-0.51	0.24	-0.39	-0.52	0.45	1.08	1.79	1.23	1.48	1.42
10	felt greasy	-0.67	0.34	-0.31	-0.42	0.20	1.86	1.61	1.27	0.91	1.23

Figure 3.2: Parallel coordinate plot of the MDS configuration before (left) and after rotation. the highlighted cluster of variables differentiates from the main group in the rotated plot.

Dim1R	1				
Dim2R	0.51	1			
Dim3R	0.52	0.58	1		
Dim4R	0.44	0.59	0.60	1	
Dim5R	0.44	0.59	0.60	0.79	1
	Dim1R	Dim2R	Dim3R	Dim4R	Dim5R

Table 3.6: The correlation between the new ordered dimensions.

3.6 Future Work

The current general approach suffers from two drawbacks. Firstly, the need to include an increasing number of extra constraints with increasing variable numbers, in order to find a unique solution. Secondly, if either criterion is constrained a solution is not guaranteed. In the case of the parallel coordinate plot this is not such a problem as the orientation of the axes is arbitrary for the representation space, and so any one of multiple solutions are potentially useful. This is also true for the general application of correlated components. However, as the aim is to find groups of components that explain different aspects of the same trait, other constraints on the optimization are desirable. For instance, to differentiate loadings and increase sparseness. In this thesis, only the general case is considered where all the covariance parameters are unconstrained. Explicit constraints on the covariance structure is a different problem and may lead to new insights.

Finding solutions that maximize the sum, or the squared sum, of the correlations of the principal components has not been considered. The scaling introduced to obtain correlations increase the complexity of finding suitable optimization criteria.

As already mentioned, explicit constraints on the covariance matrix, such as introducing zeros and block structure has not been explored.

Chapter 4

The Analysis and Utility of a Two-dimensional Response to Questions Involving Multiple Comparison

4.1 Introduction

Survey questions are traditionally scored on a number of scales. The *visual analogue scale* (VAS) tries to measure a characteristic or attitude that is believed to range across a continuum of values and cannot easily be directly measured. For example, the amount of body that a respondent perceives for their hair, after using a shampoo, ranges across a continuum from none to an extreme amount. From the respondent's perspective this spectrum appears continuous, their hair's body does not take discrete jumps, as a categorization of none, mild, moderate and high would suggest. It was to capture this idea of an underlying continuum that the VAS was devised. Operationally a VAS is usually a horizontal line, 100 mm in length, anchored by word descriptors at each end, as illustrated in Figure 4.1. The respondent marks on the line the point that they feel represents their perception of their current state. The VAS score is determined by measuring in millimetres from the left hand end of the line to the point that the respondent marks. As such an assessment is clearly highly subjective, these scales are of most value when looking at change within individuals, and are of less value for comparing across a group of individuals at one time point.

Consider the situation where a number of respondents compare product attributes but the response is now within a two dimensional VAS box. If a respondent wishes they could score an attribute along a line within the box, however they now have the option to make multiple comparisons in a more flexible way within the two dimensional box.

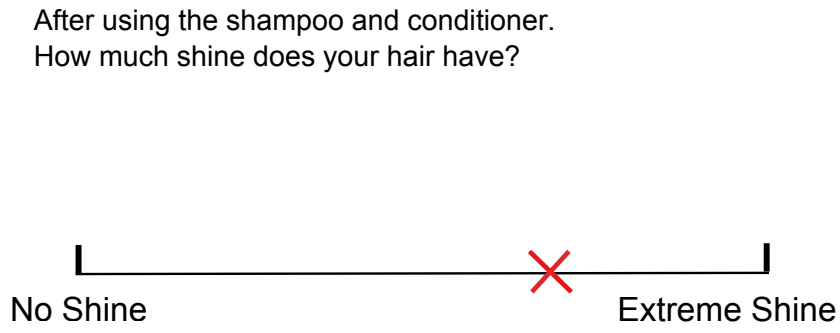


Figure 4.1: An example of the visual analogue scale

In so doing their subconscious preferences may be captured in a way that cannot by using a single line scale. These responses could take a number of forms. However, here consider a general case where the response space does not have any indication of high or low score based on orientation. Then the distances between the products indicates similarity. Figure 4.2 illustrates the ideas. The figure shows the two dimensional responses to the question *How do the products compare on softness*, for five products by three respondents.

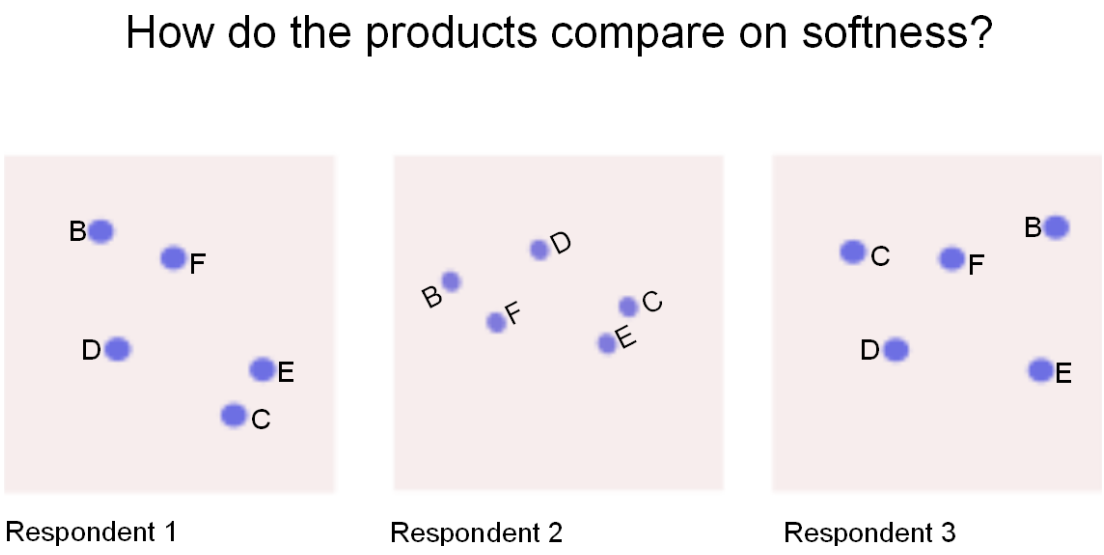


Figure 4.2: The example response of three respondents: The letters represent products compared within a five product test. The proximity of the points indicate similarity between products as assessed by each respondent. Respondent 2 can be matched onto respondent 1's response by scaling, reflection and translation and represent similar responses to the question. Respondent 3's response cannot be matched exactly onto the other two and as such is showing a different preference.

The respondents will record their response with differing notions of orientation and scale. These scale and orientation differences are nuisance parameters which need to be adjusted for. In Euclidean space such differences are covered by the group of Euclidean similarity transformations, {*translation, rotation, reflection, scale*} of the

points. The important information is the relative positions of the points to each other within a given respondent's response (ordered point set). These ordered point sets called *c*onfigurations may be considered *shapes* under a number of restrictions:

1. A point set will not generally form a shape with a closed boundary.
2. Edges, although they do not exist can be considered the Euclidean distance between points (straight lines).

These two dimensional responses may be analysed using statistical shape analysis. In the case of a set of such questions the following may be of interest,

1. Analysis of variance
2. Define models, for example a Gaussian perturbation model (Goodall, 1991, Lele, 1991)
3. Latent variable representation
4. Clustering and classification
5. Hypothesis tests

4.1.1 Toothbrush Example

Two dimensional data was collected from a small sample of respondents. Six respondents were asked to arrange seven different toothbrushes on a piece of A3 paper, in response to six questions. No indication was given as to scale or orientation. After a respondent answered a question the coordinates of each toothbrush was marked on the paper. Later the coordinates were recorded by measuring from the lower left corner of the paper to the grip of each brush. Figure 4.3 shows the raw data. Questions are across the top and respondent initials down the side, and each box is a plot of the coordinates of the toothbrushes. Each toothbrush has a unique number.

To illustrate the data, a general Procrustes superimposition was performed on each question to match the respondent's responses. Also the Procrustes mean shape was calculated for each question. The transformed data is shown in Figure 4.4 and the Procrustes mean shape in Figure 4.5. Interestingly, some respondents are using the space as a line scale, for example, PH and SB, however NM, more fully utilises the space. Looking at the Procrustes superimposition the pattern in the data becomes more apparent. For example, JK for *Healthy gums* and *Reaches difficult areas*, where the toothbrushes separate into clearer clusters. Shapes are preserved, so SB's response

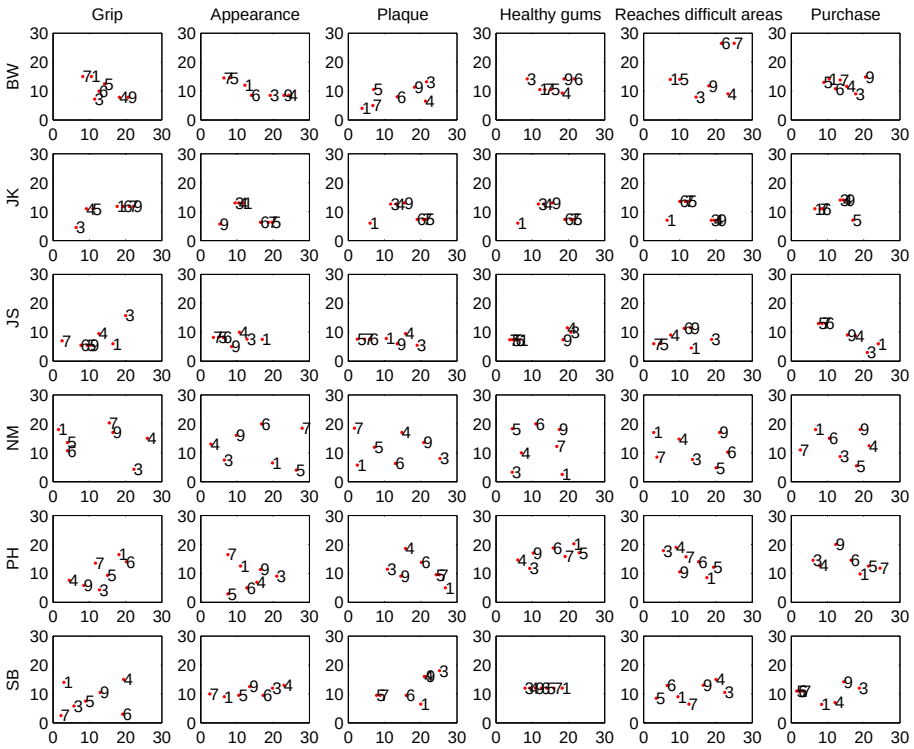


Figure 4.3: The raw toothbrush data. Each respondent arranged the seven toothbrush examples on the A3 piece of paper in response to each of the six questions. No indication of scale or orientation is presented and the respondents are free to arrange the brushes in the space as they choose.

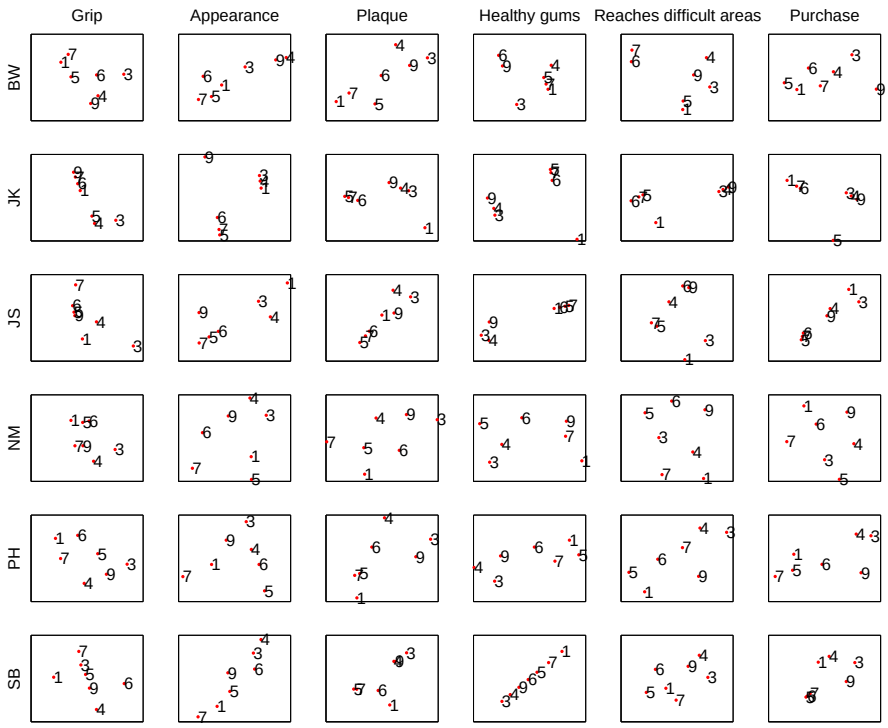


Figure 4.4: The landmark configurations for the toothbrush data after a general Procrustes superposition on each question.

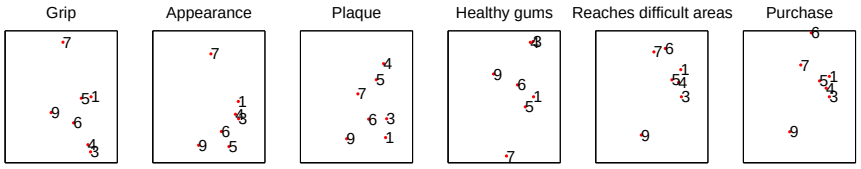


Figure 4.5: The mean Procrustes configuration for each question, across all respondents

for *Healthy gums* remains a straight line. For this small example, it appears that *Reaches difficult areas* and *Purchase* share a similar shape. Brush 7 differentiates from the others on *Grip* and *Appearance* and *Healthy gums*. Brush 9 differentiates on *Reaches difficult areas* and *Purchase*. There has been no statistical comparisons between the brushes, however, the level of *Plaque* may split the brushes into two groups.

4.2 Analysis of the Two Dimensional Response using Principal Shapes

As a useful tool for the analysis of the two dimensional VAS response, the idea of principal shapes was investigated. The motivation for principal shape analysis (PSA), stems from finding an analogue of PCA. PCA will find orthogonal axes in the data that maximise the explained variability. In the case of PCA, variability is a measure of information in the data, so this is maximized. However, with the shape data, directions where variability between shapes is high are directions of low information and increased random error. So, a direction in which the shape variability is lowest is a direction of similarity. Here, the aim is to find a formalization to extract directions of shape similarity for the two dimensional VAS responses. In the spirit of PCA, a principal shape (PS) could be a linear combination of a set of landmark configurations representing responses to questions, such that it pursues directions that minimize the principal shape variability. One difference from PCA is that a weighted sum is replaced by a weighted average,

$$PS = w_1\mathbf{X}_1 + \ldots + w_p\mathbf{X}_p,$$

where $\mathbf{X}_i$ is a landmark configuration and w is a weight, such that $\sum_i w_i = 1$ and all $w_i > 0$. The consequence of this is that the problem of finding the w_i's is no longer an eigenvalue problem.

In order to use the raw configuration data, which is a set of landmark coordinates, the configurations would be registered into a shape space coordinate system. One way to do this would be to use *general Procrustes superimposition*, and so remove translation, rotation, reflection and scale. However, Lele and Richtsmeier (2000) pointed to potential problems associated with the incorrect estimation of the variance-covariance matrix. In particular the bias introduced by constrained nuisance Euclidean transformation parameters. To avoid this a coordinate free approach is used, based on the Euclidean distance matrix, see Section 1.4. A *form* consists of the matrix of all inter-landmark distances and defines a unique point in the form space. Lele and Richtsmeier show that this enables a consistent algebra for forms to be constructed. Then,

$$F\left(PS\right) = w_1F(\mathbf{X}_1) + \ldots + w_pF(\mathbf{X}_p)$$

can be interpreted as a unique form.

4.2.1 The Variability of the Principal Shapes

The variability of a shape as defined by its Euclidean distance matrix is the sum of the squared inter landmark distances. Let $\mathbf{X}$ be a $k \times m$ matrix representing k landmarks of dimension m. The quantity D_X is the sum of the squared distances between landmarks

$$D_X = \sum_{r}^{k} \sum_{s>r}^{k} d_{rs}^2.$$

This sum is obtained from the cross product matrix $\mathbf{X}\mathbf{X}'$,

$$D_X = k \cdot \text{trace}(\mathbf{X}\mathbf{X}') - \underline{\mathbf{1}}'\mathbf{X}\mathbf{X}'\underline{\mathbf{1}}, \tag{4.1}$$

where $\underline{\mathbf{1}} = [1, \ldots, 1]'$. To show this the squared Euclidean distance between two points, $\mathbf{x}_r, \mathbf{x}_s$ is,

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s) = \mathbf{x}_r'\mathbf{x}_r + \mathbf{x}_s'\mathbf{x}_s - 2\mathbf{x}_r'\mathbf{x}_s.$$

There are $k(k-1)/2$ distances between the k landmarks. Summing these up and simplifying the indices,

$$D_X = \sum_{r}^{k} \sum_{s>r}^{k} d_{rs}^2 = k \left(\sum_{r} \mathbf{x}_r'\mathbf{x}_r - \sum_{s>r} \mathbf{x}_r'\mathbf{x}_s \right).$$

This is identical to equation (4.1) after expansion into it's sum of squares and cross-products.

A weighted average of the configurations, such as from a set of two dimensional responses, is given by

$$\mathbf{Y} = w_1\mathbf{X}_1 + w_2\mathbf{X}_2 + w_3\mathbf{X}_3 + \cdots + w_n\mathbf{X}_n,$$

where $\sum_{i=1}^{n} w_i = 1$ and $w_i > 0 \quad \forall\, i$. Then the variability D_Y of this linear combination is,

$$\begin{aligned} D_Y = k \cdot \text{trace}((w_1\mathbf{X}_1 + \cdots + w_n\mathbf{X}_n)(w_1\mathbf{X}_1 + \cdots + w_n\mathbf{X}_n)') \\ -\underline{\mathbf{1}}'(w_1\mathbf{X}_1 + \cdots + w_n\mathbf{X}_n)(w_1\mathbf{X}_1 + \cdots + w_n\mathbf{X}_n)'\underline{\mathbf{1}}. \end{aligned}$$

Using the distributive law this can be written succinctly as,

$$D_Y = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \left[k \times \text{trace}\left(\mathbf{X}_i\mathbf{X}_j'\right) - \underline{\mathbf{1}}'\mathbf{X}_i\mathbf{X}_j'\underline{\mathbf{1}} \right],$$

and

$$D_Y = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j D_{ij}.$$

Finally, written compactly in vector form,

$$D_Y = \mathbf{w}'\mathbf{D}\,\mathbf{w},$$

with $\mathbf{w} = [w_1, w_2, \ldots]'$ and $\mathbf{D} = [D_{ij}]$.

The D_{ii} elements represent the variability of the ith configuration, but what do the D_{ij} elements represent? Clearly they represent some relationship between the ith and jth configuration. These can be thought of as a measure of covariance. If the D_{ij} elements are expressed as a set of inner products by letting $\mathbf{X}_i = (x_1^i, x_2^i, \ldots, x_k^i)'$, where $x_1^i = (x_{11}, x_{12}, \ldots, x_{1m})$ is the first m dimensional landmark of the ith shape, then,

$$D_{ij} = \text{trace}[(x_1^i, x_2^i, \ldots, x_n^i)' \cdot (x_1^j, x_2^j, \ldots, x_n^j)] - \frac{1}{k}\underline{\mathbf{1}}'(x_1^i, x_2^i, \ldots, x_n^i)' \cdot (x_1^j, x_2^j, \ldots, x_n^j)\underline{\mathbf{1}}$$

and

$$D_{ij} = \text{trace}\left[\sum_{p=1}^{k}\sum_{q=1}^{k} x_p^i \cdot x_q^j\right] - \frac{1}{k}\underline{\mathbf{1}}'\sum_{p=1}^{k}\sum_{q=1}^{k} x_p^i \cdot x_q^j\underline{\mathbf{1}}$$

which reduces to

$$D_{ij} = \frac{k-1}{k}\sum_{q=1}^{k} x_q^i \cdot x_q^j - \frac{1}{k}\sum_{p=1}^{k}\sum_{q\neq p}^{k} x_p^i \cdot x_q^j.$$

The first term represents the inner products between corresponding landmarks and the second between the non-corresponding landmarks. Maximizing this expression will push corresponding landmarks between two configurations closer together and non-corresponding further apart.

4.2.2 Population Principal Shapes

Let $\mathbf{x}$ be a $p \times 1$ vector of p Euclidean shape vectors and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]$ be a matrix of weight vectors, $\mathbf{w}_i = [w_{i1}, \ldots, w_{ip}]'$. Where $\mathbf{w}_i\underline{\mathbf{1}} = 1$ and $w_{ij} > 0$. Let, $\mathbf{y}$ be a $p \times 1$ vector of principal shapes obtained from a linear transformation of $\mathbf{x}$,

$$\mathbf{y} = \mathbf{W}'\mathbf{x}$$

The covariance of the principal shapes is given by

$$\begin{aligned}
\text{cov}\left(\mathbf{y}_i, \mathbf{y}_j\right) &= \text{cov}\left(\mathbf{w}_i'\mathbf{x}, \mathbf{w}_j'\mathbf{x}\right) \\
&= \mathbf{w}_i'\text{cov}\left(\mathbf{x}, \mathbf{x}\right)\mathbf{w}_j \\
&= \mathbf{w}_i'\mathbf{D}_X\mathbf{w}_j.
\end{aligned}$$

$\mathbf{D}_X = [D_{ij}]$ and $D_{ij} = \text{cov}\,(x_i, x_j) = k \times \text{trace}\left(x_i x_j'\right) - \underline{\mathbf{1}}'\left(x_i x_j'\right)\underline{\mathbf{1}}$. The variance-covariance matrix for $\mathbf{y}$ is then

$$\text{cov}\,(\mathbf{y}) = \mathbf{W}'\mathbf{D}_X\mathbf{W}$$

4.2.3 Sample Principal Shapes

In the case of a PCA the raw data is centred or standardized. If the raw configurations are used then the analogy here is the removal of the Euclidean nuisance parameters of translation, rotation, reflection and scale using a general Procrustes superimposition. However, if the Euclidean shape matrix is used then there is no requirement (other than the scaling of the Euclidean form to give a shape). For a set of m respondents and p questions a configuration data matrix $\mathbf{Z}$ can be constructed, where $\mathbf{s}_{ij}$ is the ith respondent's Euclidean shape distance vector for the jth question.

$$\mathbf{Z} = \begin{pmatrix} \mathbf{s}_{11} & \dots & \mathbf{s}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{s}_{m1} & \dots & \mathbf{s}_{mp} \end{pmatrix}.$$

The sample principal shape scores $\mathbf{Y}$ are then,

$$\mathbf{Y} = \mathbf{Z}\mathbf{W},$$

where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_p)$ is a matrix whose columns consist of the weight vectors, and $\mathbf{w}_i'\underline{\mathbf{1}} = 1$.

A sample covariance matrix can be constructed from this data. The covariance between two shapes is given by

$$\text{cov}\,(\mathbf{s}_1,\ \mathbf{s}_2) = \mathbf{s}_1'\ \mathbf{s}_2.$$

The sample covariance matrix for the questions is then the covariance averaged over the respondents. The estimate of each covariance parameter is

$$\hat{D}_{ij} = \frac{1}{N-1}\sum_{n=1}^{N} \mathbf{s}_{in}'\mathbf{s}_{jn},$$

where $\mathbf{s}_{in}$ is the shape for the nth respondent for the ith question.

Then the covariance matrix is

$$\mathbf{D}_Z = \text{cov}\,(\mathbf{Z}) = \frac{1}{N-1}\mathbf{Z}'\mathbf{Z}$$

where

$$\mathbf{D}_Z = \frac{1}{m-1}\begin{pmatrix} \sum_{k=1}^{m} D_{11}^k & \sum_{k=1}^{m} D_{12}^k & \dots & \sum_{i=1}^{m} D_{1p}^k \\ \sum_{k=1}^{m} D_{12}^k & & & \\ \vdots & \ddots & & \\ \sum_{k=1}^{m} D_{1p}^k & & & \sum_{k=1}^{m} D_{pp}^k \end{pmatrix},$$

and D_{ij}^k is the covariance between question i and j for the kth individual.

4.2.4 The Questionnaire Framework

A framework for analysing the two dimensional response data could be as follows. A series of Q questions relating to K objects are presented to each of N respondents. The respondents are asked to arrange the objects on a suitably sized sheet of paper in such a way that the arrangement represents their perception of how the objects compare in response to the current question. No indication of orientation or scale is given. The respondents could be additionally asked to indicate on the paper where they perceive high and low anchor points to be. However, this is not considered in the first instance. The position of each object is marked on the sheet with their identifier before continuing to the next question.

Then the data set consists of a $N \times Q$ matrix of landmark configurations of dimension $K \times 2$. Each configuration is converted to a column vector of ordered Euclidean inter-landmark distances to give its form, denoted as f_{nq}. As subjects may scale their responses differently, the forms are normalised using a measure of size. In the case of a form this can be the geometric mean of the inter-landmark distances,

$$S(f) = \{\prod d_i\}^{1/L} \quad i = 1, \ldots, L.$$

L is the number of inter-landmark distances (dimension of the form space) and is given by

$$L = \frac{K(K-1)}{2}.$$

4.2.5 Finding Principal Shapes

The set of weight vectors are found that sequentially minimize the covariance matrix of the principal shape scores (produce a set of principal shapes which are most similar). Consequently, the later principal shapes will have higher variation associated with them, and so represent non-conformity between the respondent's responses to the questions. There is no concept of a negative shape, and so a principal shape is a weighted average of the sample shapes. The subsequent weighted averages represent the principal shapes within the population. The first principal shape is found by minimizing,

$$\mathbf{w}_1' \mathbf{D}_Z \mathbf{w}_1,$$

where $\mathbf{w}_1$ is a weight vector. This is similar to PCA, however there are additional constraints. The coefficients of the weight vector must be positive and sum to one, The

second principal shape is found subject to the first. The second weight vector cannot be constrained to be orthogonal to $\mathbf{w}_1$, as this is infeasible due to the previously mentioned constraints. A penalty parameter is used to push the direction of $\mathbf{w}_2$ as far away from $\mathbf{w}_1$ as possible. The third principal shape is found subject to the first and second and so on.

In general the following is minimized,

$$L(\mathbf{w}_i, \boldsymbol{\mu}) = \mathbf{w}_i' \mathbf{D}_Z \mathbf{w}_i + \tau \sum_{j=1}^{i-1} \left(\mathbf{w}_i' \mathbf{w}_j\right)^2 + \sum_{j=1}^{i} \mu_j \left(\mathbf{w}_j \underline{\mathbf{1}} - 1\right) \quad i = 1, \ldots, Q, \qquad (4.2)$$

where the weight vector elements $w_{ij} \geq 0 \; \forall \; ij$, $\sum_j w_{ij} = 1 \; \forall \; i$ and μ_j is a Lagrange multiplier and τ is a penalty parameter.

This is a *standard quadratic optimization problem* (see, Fletcher, 2000). The principal shapes can be converted into a representative landmark configuration (icon) by applying metric scaling, see Section 1.4.3.

4.3 Principal Shape Analysis of the Toothbrush Data

The analysis was performed using the shape representation based on the Euclidean distance matrix (EDM). The resulting covariance matrix was standardized by scaling so that the variances were unity. The resulting correlation matrix was positive definite (as was the covariance matrix). A range of values for the penalty parameter were experimented with. These gave slightly sparser solutions as the penalty vector increased. The range used was {1, 10, 100, 1000}, but in all the cases the results were similar. The use of the correlation matrix set the total variance in the data to six. Consequently, a value of ten was chosen for the penalty as this was similar in magnitude. In order to visualize the resulting principal shapes, the respondent's principal shape scores were averaged to give six mean principal shapes. The mean principal shapes were then converted from their EDM representation to icons in two dimensional coordinate space using metric scaling. Finally to aid visual interpretation they were matched using general Procrustes superimposition. Figure 4.6 shows the mean principal shapes and Table 4.1 shows the weight vectors. The vectors do not represent orthogonal axes as this is infeasible. Also their lengths are not constrained to be unity.

Visual examination of the mean principal shapes suggests that respondents split the toothbrushes into distinct groups. Brushes 4 and 9 differentiate on all the mean principal shapes. Brush 3 is grouped with 4 and 9 except in PS6. PS1 and PS2 position the brushes similarly, except 5 and 6 are transposed. In PS3, 5 and 7 are similar, however,

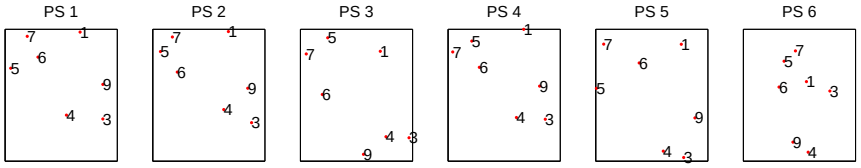


Figure 4.6: The mean principal shapes for the toothbrush data.

	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_6$	
Grip	0.32	0.37	0.20	0.06	0.00	0.19	
Appearance	0.20	0.00	0.08	0.00	0.71	0.04	
Plaque	0.00	0.46	0.00	0.35	0.29	0.00	
Healthy gums	0.09	0.00	0.00	0.21	0.00	0.71	
Reaches difficult place	0.00	0.00	0.46	0.38	0.00	0.00	
Purchase	0.39	0.17	0.26	0.00	0.00	0.05	
							total
Variance explained	0.87	0.89	0.90	0.92	0.94	0.93	5.44

Table 4.1: The weight vectors for the mean principal shapes. All individual weights are positive and the vectors sum to one.

6 and 1 are separate from each other and the two groups. Toothbrush 1 is consistently separate from the other brushes, except on PS6. Examining the weight vectors, $\mathbf{w}_1$ is an average of *Grip, Appearance* and *Purchase*. This represents the principal shape for which most respondents agree, i.e. lowest variability. As mentioned this is characterized by three clusters of brushes, (5,6,7), (3,4,9) and 1. There is no indication as to which were preferred. The vector $\mathbf{w}_2$ are the weights for PS2 and represents the principal shape that was most similar between respondents, after PS1. This is characterized by *Grip* and *Plaque*; also *Purchase*, but to a lesser degree. PS2 is characterized by the same toothbrush groups as PS1. In fact the principal shape patterns are very similar for PS1, PS2 and PS3, but they appear to extract different aspects of a respondents propensity to purchase a toothbrush. PS3 is loaded heaviest on *Reaches difficult places*. PS4 could be interpreted as a mouth health principal shape. PS5 and PS6 which represent the highest disagreement between respondents differentiate heavily on *Appearance* and *Healthy gums* respectively. Interestingly, 4 and 9 are set apart on PS6 , which is dominated by *Healthy gums*.

4.4 Future and Related Work

In addition to finding principal shapes, it is interesting to estimate the variability of particular toothbrushes (landmarks). Using a EDM gives consistent estimates of this variability in the population (see section 1.4.3). Also, the objective function of equation (4.2) is currently solved sequentially, but weight vectors could be found simul-

taneously.

Another approach that was considered is the use of individual difference scaling (INDSCAL) (Cox and Cox, 2000) to represent the respondents and questions on separate two dimensional maps. However, it was discovered that this approach is available using a technique called *napping*, see Perrin et al. (2008) for example, which analyses the napping data using hierarchical multiple factor analysis. Packages are available in R, to use napping data with INDSCAL, and is part of the SensoMineR package (Le and Husson, 2008).

Bibliography

Anaya-Izquierdo, K., Critchley F. and Vines K. (2008). Orthogonal simple component analysis. *Vectors 1*, z4.

Bartholomew, D. J. and Knott M. (1987). *Latent variable models and factor analysis*. Griffin London.

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*. Wiley & Sons.

Berge, J. M. T. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika 42*(2), 267–276.

Bookstein, F. L. (1984). A statistical method for biological shape comparison. *Journal of Theoretical Biology 107*(3), 475–520.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science 1*(2), 181–242.

Brusco, M., Hubert L., Arabie P. and Meulman J. (2007). The structural representation of proximity matrices with MATLAB. *Psychometrika 72*(4), 655–656.

Burt, C. L. (1917). *The distribution and relations of educational abilities*. London County Council.

Cadima, J. and Jolliffe I. T. (1995). Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics 22*(2), 203.

Cadima, J. F. C. L. and Jolliffe I. T. (2001). Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and Environmental Statistics 6*(1), 62–79.

Chipman, H. A. and Gu H. (2005). Interpretable dimension reduction. *Journal of Applied Statistics 32*(9), 969.

Choulakian, V. (2001, August). Robust q-mode principal component analysis in l1. *Computational Statistics & Data Analysis 37*(2), 135–150.

Choulakian, V. (2003). The optimality of the centroid method. *Psychometrika 68*(3), 473–475.

Choulakian, V. (2005, January). Transposition invariant principal component analysis in l1 for long tailed data. *Statistics & Probability Letters 71*(1), 23–31.

Choulakian, V. (2006a, March). L1-norm projection pursuit principal component analysis. *Computational Statistics & Data Analysis 50*(6), 1441–1451.

Choulakian, V., Allard J. and J. Almhana (2006b). Robust centroid method. *Comput. Stat. Data Anal. 51*(2), 737–746.

Choulakian, V., Dambra L. and Simonetti B. (2006c). Hausman principal component analysis. *From Data and Information Analysis to Knowledge Engineering*, 294–301.

Cootes, T. F., Taylor C. J., Cooper D. H. and J. Graham (1992). Training models of shape from sets of examples. *557*, 9–18.

Cormen, T. H., Leiserson C. E., Rivest R. L., and Stein C. (2001). *Introduction to Algorithms, chapter 16.* Cambridge (Massachusetts): MIT Press.

Costello, A. B. and Osborne J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation 10*(7), 19.

Cox, T. F. (2005). *An introduction to multivariate data analysis.* Hodder Arnold London.

Cox, T. F. and Cox M. A. A. (2000). *Multidimensional Scaling.* Chapman and Hall.

D'Aspremont, A., Ghaoui L. E, Jordan M. I. and Lanckriet G. R. (2004). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review 49*(3), 434–448.

DeSarbo, W. S. and Hausman R. E. (2005). An efficient branch and bound procedure for restricted principal components analysis. *Data Analysis and Decision Support*, 11–20.

Dryden, I. L. and Mardia K. V. (1998). *Statistical Shape Analysis.* Wiley.

Edwards, D. (2000). *Introduction to Graphical Modelling.* Springer-Verlag, New York.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models.* Monographs on Statistics and Applied Probability. Chapman & Hall.

Fabrigar, L. R., Wegener D. T., MacCallum R. C. and Strahan E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods 4*(3), 272–299.

Ferrez, J. A., Fukuda K., and Liebling T. M. (2005). Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm. *European Journal of Operational Research 166*(1), 35–50.

Fletcher, R. (2000). *Practical methods of Optimization* (2nd ed.). Wiley.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika 58*(3), 453.

Gervini, D. and Rousson V. (2004). Criteria for evaluating Dimension-Reducing components for multivariate data. *The American Statistician 58*(1), 72–77.

Goodall, C. R. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society, Series B 53*(2), 285.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrica 40*(1), 33–51.

Gower, J. C. and Hand D. J. (1996). *Biplots* (First ed.). Monographs on Statistics and Applied Probability. Chapman & Hall.

Gragn, D. and Trendafilov N. T (2010). Sparse principal components based on clustering. Technical report, The Open University.

Harman, H. H. (1976). *Modern factor Analysis*. Chicago University Press.

Harris, C. W. and Kaiser H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika 29*(4), 347–362.

Hausman, R. (1982). Constrained multivariate analysis. *Optimization in Statistics*, 137.

Hendrickson, A. E. and White P. O. (1964). Promax: a quick method for the rotation to oblique simple structure. *British Journal of Mathematical Statistics and Psychology 17*(1), 65–70.

Jennrich, R. (2002, March). A simple general method for oblique rotation. *Psychometrika 67*(1), 7–19.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions (with discussion). *Journal of the American Statistical Association 104*, 682–703.

Jolliffe I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics 22*(1), 29-35.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer verlag.

Jolliffe, I. T., Trendafilov N. T. and Uddin M. (2003, September). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Jolliffe, I. T. and Uddin M. (2000). The simplified component technique: An alternative to rotated principal components. *Journal of Computational and Graphical Statistics 9*(4), 689–710.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187–200.

Kalman, D. (2000, October). A matrix proof of newton's identities. *Mathematics Magazine 73*(4), 313–315.

Kendall, D. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin London Mathematical Society 16*, 81.

Kent, J. T. and Mardia K. V. (1997). Consistency of procrustes estimators. *Journal of the Royal Statistical Society, Series B 59*(1), 281–29.

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin: Springer.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika 29*(2), 115–129.

Le, S. and Husson F. C. (2008). SensomineR: a package for sensory data analysis. *Journal of sensory studies 23*(1), 14–25.

Lele, S. (1991). Euclidean distance matrix analysis:a coordinate-free approach for comparing biological shapes using landmark data. *Am. J. Phys. Anthropol 86*, 415.

Lele, S. (1993). Euclidean distance matrix analysis (EDMA) of landmark data:estimation of mean form and mean form difference. *Math. Geol. 25*(5), 573–602.

Lele, S. and Cole T. (1996). A new test for shape differences when variance-covariance matrices are unequal *Journal of Human Evolution 31*, 193.

Lele, S. R. and Richtsmeier, J. T. (1992). *On comparing biological shapes: detection of influential landmarks. Am. J. Phys. Anthropol 87*, 49.

Lele, S. R. and Richtsmeier, J. T. (1995). *Euclidean Distance Matrix Analysis: Confidence intervals for form and growth differences. Am. J. Phys. Anthropol 98*, 73–86.

Lele, S. R. and Richtsmeier J. T. (2000). *An Invariant Approach to Statistical Analysis of Shape*. Interdisciplinary Statistics series. London: Chapman and Hall/CRC Press.

MacKay, D. J. and Gibbs M. N. (1999). Density networks. *Statistics and neural networks: advances at the interface*, 129.

Mardia, K. V., Kent J. T. and Bibby J. M (1979). *Multivariate Analysis*. London:Academic Press.

Mead, D. G. (1992, October). Newton's identities. *The American Mathematical Monthly 99*(8), 749–751.

Mosier, C. I. (1939). Determining a simple structure when loadings for certain tests are known. *Psychometrika 4*(2), 149–162.

O'Neill, B. (1997). *Elementary differential geometry*. Academic Pr.

Perrin, L., Symoneaux R., Matre I., Asselin C., Jourjona F. and Pagesc J. (2008). Comparison of three sensory methods for use with the napping procedure: Case of ten wines from loire valley. *Food Quality and Preference 19*(1), 1–11.

Rousson, V. and Gasser T. (2004). Simple component analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 53*(4), 539–555.

Rousson, V. and Maechler M. (2009, November). R package 'sca' - simple component analysis.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers 100*(18), 401–409.

Slabaugh, G. G. (1999) Computing Euler angles from a rotation matrix. Report. http://gregslabaugh.name/publications/euler.pdf

Svensen, J. F. M. (1998). *GTM:The Generative Topographic Mapping*. Ph. D. thesis, Aston University.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review 38*(5), 406–427.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Vigneau, E. and Qannari E. M. (2003). Clustering of variables around latent components. *Communications in statistics. Simulation and computation 32*(4), 1131–1150.

Vines, S. K. (2000). Simple principal components. *Applied Statistics 49*(4), 441–451.

Watson, G. S. (1986). The shape of a random sequence of triangles. *Advances in Applied Probability 18*(1), 156–169.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Witten, D. M., Hastie T. and Tibshirani R. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Zou, H., Hastie T. and Tibshirani R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.